Bauhaus-Universität Weimar
Faculty of Media
Degree Programme Computer Science for Digital Media

Retracing the travel path of Marco Polo

Master's Thesis

Ruta Hareshbhai Bakhda

1. Referee: Prof. Dr. Benno Stein
2. Referee: Jun.-Prof. Dr. Jan Ehlers

Submission date: April 8, 2021

Declaration

Unless otherwise indicated in the text or references, this thesis is entirely the product of my own scholarly work.

Weimar, April 8, 2021

.....
Ruta Hareshbhai Bakhda

Abstract

The 12th century travelogue *The travels of Marco Polo* holds an important place in human history and this thesis proposes a method to reconstruct the travel path of *Marco Polo* from the digitized narrative using Natural Language Processing techniques. Marco Polo visited numerous places throughout his journey and these place references are identified using state-of-the-art Named Entity Recognition techniques and domain-specific gazetteer created in this thesis. To separate the stories related to travel from the other stories in a narrative, motion events are extracted using various lexical resources. Later, extracted motion event triggers and extracted places references are linked to identify locations visited by *Marco Polo*. These locations are then mapped to their contemporary equivalents and placed on a map to generate a travel path. The quality of generated path and other Natural Language Processing tasks is evaluated using the gold standard annotations generated in this thesis in terms of precision, recall and F1 score. This thesis also provides in-depth quantitative and qualitative analysis of the narrative and discusses the problems associated with various Natural Language Processing techniques when applied to historical corpora. Results for path retracing shows that travelled location that are explicitly described as being visited can be found but traveled locations that are not explicitly mentioned as visited in the narrative are not identified using motion events and it needs further research.

Contents

Chapter 1

Introduction

1.1 Thesis Context

An increasing number of historical texts are becoming available in digital form and it has led to a growth in interest in applying Natural Language Processing (NLP) techniques to a variety of historical texts that can provide useful information for the study of these historical texts. The 12th century narrative named *The travels of Marco Polo* has shaped the history by establishing ties between Europe and China. This narrative has been an important resource for mapmakers since the year it was written.

A significant amount of domain knowledge is essential for extracting information from the narrative texts. However, this information extraction is mostly a manual task and requires a human interpretation, a notoriously time-consuming task for which expertise in knowledge engineering is essential. Now if we can automatically extract various narrative information from the natural language, we can leverage the vast amount of content from the written literature. However, this presents significant new challenges, especially for historical corpora as old as 12th century which has quite a different writing style.

The goal of this thesis is to separate the stories related to the path taken by *Marco Polo* from other stories in the narrative and to identify the locations he actually visited. The other stories include *Marco Polo's* experience about different cultures and kingdoms he encountered on his way, their customs, ways of living, their trades and war stories. These different stories can be separated by identifying events and focusing only on motion events to find the travel path. In addition, various place references needs to be identified to find the locations visited by *Marco Polo*. However, in the case of historical documents, there are several issues such as a) identification of place references and b) place reference ambiguity. Issue a) can present challenges related to Optical Character Recognition errors, language changes over time, spelling variations,

non-existing locations. Where as issue b) involves challenges related to correct mapping of place descriptions from the text to a particular place. In Example 1.1, Marco Polo describes a province and its capital, both sharing the same name. Also, a naming convention uses here special character, like in a place name *Sin-din-fu*. This naming convention is a very common throughout the book but it is not prevalent today and hence, it creates significant challenges for the modern Natural Language Processing tools. Here, the ability to detect and project place names with reasonable accuracy is very crucial.

(Example 1.1) *"When a man has left this country and traveled twenty days westward, he approaches a province on the borders of Manji named Sin-din-fu. The capital, bearing the same name, was anciently very great and noble, governed by a mighty and wealthy sovereign."*

Another observation to note is, there are several unnamed place references throughout the entire book which makes it very difficult to determine the distance between two referenced places. As shown in Example 1.2, this sentence gives information about the source location (Sin-gui) from which Marco Polo goes to some other destination location. However, this sentence also has some other information, such as direction (south) and duration (eight days) and several unnamed place references (many cities and castles). Now without knowing the mode of transportation and the time he spent in the traveling each day, it is very difficult to determine how much distance Marco Polo covered in eight days and hence, where he would have reached in the direction of south at the end of eight days.

(Example 1.2) *"When a man departs from Sin-gui and goes eight days to the south, he finds many cities and castles."*

In addition, *Marco Polo* talks about his experiences and locations he visited in a chronological order as he continues his journey so extracted locations can be connected in a sequence to find the travel path of *Marco Polo*. This is verified by the author of the thesis when the narrative is examined manually to extract traveled locations and when these manually extracted locations were connected in a sequence to generate a travel path. However, for the complete construction of the journey of *Marco Polo*, i.e, starting the journey from Venice, traveling through middle-east, exploring China and journey back to Venice via Indian subcontinent, it requires additional domain knowledge of changing the sequence of three parts of the book. The reason behind that is, the focus of the narrative is Polo's exploration in China and his experiences at the court of

Kublai Khan and hence, they are described in the first part of the book. The second part describes the beginning of the journey from Venice to China via middle east. So through manual examination, the orders of Part One and Part Two are interchanged. However, the order of travel path within any individual part of the narrative is in the chronological order of travel. In addition, there is a difference of opinion among researchers concerning the exact travel path taken by *Marco Polo* as there are several named and unnamed place references without explicitly stating whether Marco Polo actually visited them or he is describing the places he heard about from the people during his journey. Hence, the path identified in this thesis is compared against the gold standard travel path generated by the author of the thesis.

Now also, the place descriptions can be found in a variety of contexts in a narrative such as describing a country or a city, describing the people or describing the specialty of certain regions, describing war stories between two regions, etc. Hence, it is important to find place reference mentions only associated with events related to motion or travel. It can be achieved by identifying motion events and then, establishing a link between identified place descriptions and motion events.

1.2 Research Objectives

The goal of this thesis is to examine the possibility of retracing the path of travel from the narrative of *Marco Polo*. Due to the arbitrary nature of the natural language in this historical corpora that contains the narratives of various different characters, it will be a challenging task to separate the travel path of *Marco Polo* from the rest of the narrative. Also, current-state-of-the-art methods in the field of Natural Language Processing will be reviewed and evaluated against this objective and an approach will be proposed to find place references visited by *Marco Polo* and to link them into a path. Three main objectives will be the center of an investigation:

1. Evaluating Named Entity Recognition (NER) tools for the identification of place names in historical corpora

2. Motion events identification to separate stories of travel from the other stories in narrative

3. Linking motion event and the place references to find traveled place references

1.3 Course of Investigation

This thesis is structured as follows:

Chapter 2 presents an overview of the background knowledge required for this research.

Chapter 3 presents the resources and the benchmark tools involved in this research.

Chapter 4 represents corpus, its quantitative analysis and the phenomena observed in the dataset. It also discusses the challenges presented by the historical dataset for the Natural Language Processing and also discusses the limitation and ambiguity of the investigated dataset. It also describes the gold standard setup to measure the quality of the travel path generated.

Chapter 5 describes the methodology implemented to construct a travel path from the narrative text.

Chapter 6 discusses the quality of the Named Entity Recognition tools in terms of historical data and the quality of motion event identification for finding locative expressions along with the quality of the travel path generated.

Chapter 7 summarizes the research work done in this thesis and the success of the suggested approach along with describing the ideas for the future work.

Chapter 2

Background

This chapter serves as an introduction to research work of Natural Language Processing techniques that are involved in this thesis to extract the travel path from the historical narrative. Two Natural Language processing tasks, Named Entity Recognition and Motion Event Extraction, which play a crucial role in this thesis, are described. Section 2.1 deals with background research in terms of Named Entity Recognition for the historical dataset. Section 2.2 deals with event extraction to find motion event triggers. In addition, it also describes the method of shallow parsing to find locative expressions that are linked to motion events. These locative expressions contain location entities and additional information about those location entities. These locations entities from locative expressions can be connected in a chronological order to generate a travel path.

In the past several years with the advancement in technology, there has been the massive digitization of the textual resources [17]. Hence, there is an increasing interest in applying Natural Language Processing techniques to historical texts [45] for various applications in the fields of digital humanities [23], information retrieval [9], historical linguistics [34]. However, when NLP techniques are applied to historical research, several methodological issues arise [15] because these historical texts differ from the modern text in a number of linguistic aspects such as spelling variations [19], linguistic variations [7], and syntax structures [12].

For the task of a path extraction from the historical text, several attempts have been made by various researchers. One of the studies by Barbaresi that is focused on a historical text, retraces the path taken by the author of the historical text in the Shandong province by combining coordinates, sequence and sense of time [2] on Richthofens Travel Journals from China (1907) and Die Fackel magazine (1899-1936). However, the 12th century travelogue of *Marco Polo* contains many extinct locations which have gone under significant

changes or they do not longer exist. It is almost near to impossible to locate these extinct locations on a current-day map and to find coordinates. Hence, this thesis will focus only on the sequence to retrace the path taken by *Marco Polo*.

2.1 NER in Historical Corpora

Named Entity Recognition (NER) is the subfield of Natural Language Processing that uses machine learning techniques and aims at identifying and classifying named entities in unstructured texts by identifying words and classifying them into predefined categories such as persons, locations and organizations among others [40] and it is one of the challenging problem in NLP [48]. Below is the example of NER where various entities are marked with their respective entity type.

[PERSON Marco Polo] noted that, [COUNTRY Armenia the Greater] is a large country and at the entrance of it is a city called [CITY Arzinga].

Named Entity Recognition has also become a point of interest for unstructured texts and literature because of its usefulness in information extraction. Even though Named Entity Recognition technologies have demonstrated impressive results with modern corpora, they pose a significant challenges for historical corpora and domain-specific corpora as demonstrated by several examples of research projects using Named Entity Recognition tools for historical dataset [21, 38, 59]. These projects have evaluated current state-of-the-art Named Entity Recognition systems for historical corpora along with addressing most pressing challenges posed by historical text. The languages are continuously evolving and the texts are being translated from one language to another and hence, there are issues of text containing spelling variations [4], dissimilar quality of digitization and Optical Character Recognition [21]. In addition, input data can be highly noisy with errors such as transcription errors or misspellings, and for those issues, adapted approaches have already been devised [10, 51]. Further, texts in the historical corpora often contain rare entities which do not longer exists or have gone under significant changes [57]. Also, historical corpora uses different naming convention than modern-day languages [6, 8].

Although there is an increase of interest in such research, the number of case studies dealing with historical datasets is still not significant. There have been a few attempts to evaluate different Named Entity Recognition systems for the modern historical material (19th and 20th century newspapers) [11]

but still that number is significantly low. To date, there have not been many attempts to evaluate different Named Entity Recognition systems for the 12th century old historical corpora. This thesis targets to evaluate Named Entity Recognition tools against the 12th century historical narrative of *Marco Polo*.

Apart from state-of-the-art Named Entity Recognition tools, another commonly used approach is known as Gazetteer and in that, entity references are matched against entries in a gazetteer. Gazetteer can be defined as a directory which organizes knowledge and details about place references [20]. The research work by Paradesi made an attempt to combine the Named Entity Recognition system with an external gazetteer and this system is known as TwitterTagger [42]. This system first assigns part-of-speech (POS) tags to words in a tweet text to detect proper nouns and then compares these proper nouns to the USGS database (U.S. Board on Geographic Names) [43] of locations to classify nouns that are geographic references. Also, for the cartographic visualisation, it is necessary to add geographic coordinates to a place name which is known as Geocoding and that process mostly relies on gazetteers. Gazetteers play an important role for historical research but existing digital gazetteers do not meet the challenges of historical texts [53]. In addition, historical gazetteers are not available for historical text and even though, those who are available, are available for text as late as 20th century Europe [46], and in addition, development of gazetteer is challenging as well [53]. There are some existing toolboxes, such as HeidelPlace [50]. It includes a generic gazetteer model that is an extensible framework and allows to embed information from the different knowledge bases but the use of it on historical corpora is not straightforward because of the need for additional engineering decisions to add heterogeneous gazetteer sources. Another toolkit developed by Barbaresi developed is adaptable to various historical contexts and it features bootstrapping options, geocoding and disambiguation algorithms, and cartographic processing [3] In this thesis, state-of-the-art Named Entity Recognition tools and generated domain-specific gazetteer are applied and evaluated for place references extraction.

2.2 Locative Expressions

2.2.1 Definition of Locative Expression

The main focus of this thesis is to find the locations visited by *Marco Polo* in his journey which includes automatic identification of locative expressions through motion event identification from the narrative. A locative expression (LE) is an expression that involves preposition, its object and it geolocates an entity in the text [24, 41, 62]. It identifies the place reference, the type

of entity and the relational word such as preposition. For example, "city of Maabar" is a locative expression where "city" is the pronoun that refers to the entity, the geographical reference (entity) is "Maabar" and the relational word is "of". The entities generally take a form of a noun phrase.

2.2.2 Identification of Locative Expressions

Earlier work for the identification of locative expressions focused on the detection of application specific geospatial references from the specific descriptions [58]. The work of Liu and others [35] focused on the automatic identification of locative expressions from unrestricted natural language text of social media posts and it carries out extensive error analysis to suggest ways of improving the accuracy of geoparsers. However, all the research to find locative expressions focus on the modern day language and there has been less research done for the identification of locative expression from the historical texts compare to modern texts.

There are several steps involved in identification of locative expressions such as Part-of-Speech tagging, motion event identification, shallow parsing and these are described as below. Part-of-Speech tagging helps to identify the type of the token, motion events identification technique uses these Part-of-Speech tags to identify motion event triggers and shallow parsing identifies noun phrase chunks related to place references.

Part-of-Speech Tagging

A Part-of-Speech tagging is the process of labelling each word of a sentence with its appropriate part of speech such as nouns, verbs, adjectives, adverbs etc [58]. Part-of-Speech tagging is the first step of many other complex processing of Natural Language Processing techniques such as shallow parsing. In the context of this thesis, it is further used in the identification of named entities and motion event triggers. However, lack of standardization in historical texts pose a significant challenge for modern Part-of-Speech taggers. The accuracy of Part-of-Speech software CLAWS [16] dropped from 96%-97% on written text for the British National Corpus to 82% for Early Modern text (the Shakespeare corpus) [49] affected by spelling variants. One of the research [52] improves the quality of Part-of-Speech tagging by translating historical text to equivalent modern texts. The work of Yang [61] shows that the combination of feature embedding method for domain adaption along with spelling canonicalization improves tagging accuracy by 5% for Early Modern English text. Another research by Hardmeier [22] shows that the need for spelling normalization can be eliminated by using character-level neural network to build

Part-of-Speech tagger for historical texts and that achieves results very close to the state-of-the-art solution. Here is an example of Part-of-Speech tagging where each token in a sentence is assigned Part-of-Speech tag.

[DT The] [JJ great] [NNP Khan] [VBD decided] [TO to] [VB punish] [PRP them].

Appendix A shows the most frequent Part-of-Speech tags used in Penn Treebank corpus [37].

Motion Event Identification

A verb is simply a word that shows action or a state of being. An event is an output of an action denoted by verb and hence, verb can be considered as an event trigger. Computational verb lexicon are important in Natural Language Processing for semantic interpretation of an unstructured texts. To identify motion event triggers, it is important to identify verbs that reflect the action of motion. In the field of Information Extraction, deeper knowledge structures can be exploited through a relationship between verb and its arguments denoted by predicate-argument structure. VerbNet [29], FrameNet [1] and PropBank [27, 28], these three lexical resources, have constructed the definitions for the predicate-argument frames. Predicate-argument structure analysiss key component is frames for predicates (verbs) and the roles of their arguments (part of the sentences around it). Researchers use any of these resources to extract information from the text. In this thesis, WordNet, VerbNet and FrameNet are used for the task of identification of motion event triggers.

VerbNet [29] is one of the largest lexical resource based on Levin's verb classification [33] that organizes verbs into classes and members of a class share core semantic and syntactic coherence. Each verb class in VerbNet is described by thematic roles, selection restrictions on the arguments, frames consisting of a syntactic description and semantic predicates with a temporal function. Another lexical resource is the Berkeley FrameNet project[1] which has built a lexicon that is based on the theory of Frame Semantics. Huang et al. [26] considered all verbal and lexical units in FrameNet as candidate event triggers and he defined candidate arguments by regarding all the concepts having semantic relations to candidate trigger. After that, to identify the final trigger and its arguments, similarity is calculated between each pair of triggers and arguments. Event instances are combined into different event groups each with latent yet distinguished topic based on these extracted triggers and arguments.

Another approach is Semantic Role Labeling [18, 47] that identifies different features of a sentence and these features help to extract events from the

text. One of the research work [13] used semantic role labeling along with VerbNet thematic roles for event identification in the 10% of English sample of Wikipedia articles and it has misidentified agents as a frequent source of errors. Another approach to identify event triggers is presented in this research [30] where they developed a SVM based supervised system in conjunction with various techniques based on Semantic role labeling, WordNet and handcrafted rules for event extraction. However, it requires further work of defining more precise rules for event identification and also for multi-word events.

However, there is less research available that attempts at identifying events from historical texts. One of the research [54] has released new annotation guidelines and models for automatic annotation of event mentions and types, and it has categorized events into 22 classes.

Shallow Parsing

Shallow parsing is widely used in many language processing tasks [39]. Structurally analyzing text allows extracting constituents of it such as nouns, verbs, noun groups, verb groups etc. It is also known as chunking or light parsing and also it does not help to identify internal structure of a sentence. In this process, sentence is broke down into a several non-overlapping various chunks and each chunk is assigned one of out of several labels as shown in Appendix B. When a sentence is shallow parsed, it is possible to identify verbs phrases containing motion event triggers and noun phrases containing extracted location entities. Here is the example of shallow parsing where each chunk is surrounded by parenthesis pair[ ] where NP represents Noun Phrase, PP represents Prepositional Phrase and VP represents Verb Phrase.

[NP Khan] [VP is] [PP in] [NP his Palace] [PP at] [NP Kambalu].

Now there can be more than one phrases of any of these types in a sentence. Once these phrases have been identified, a relationship between each verb phrase containing motion event trigger and each noun phrase containing location can be examined to identify whether they are linked and if thry are linked, then it can be concluded that the extracted location has been visited by *Marco Polo*.

Chapter 3

Resources

This chapter describes the resources used in this thesis. Section 3.1 describes the Named Entity Recognition tools used in the thesis for the identification of location entities and Section 3.2 describes the lexical resources for English language which represents the semantic similarities between verbs and its arguments. This similarity is used to identify the motion event and to identify a link between motion events and location entities.

3.1 Named Entity Recognition Tools

Stanford NER

Stanford NER[1] is based on supervised learning which uses linear chain Conditional Random Field sequence models and hence, it is also know as CRFClassifier [14]. The linear chain CRF [31] is a probabilistic discriminative model that uses the contextual information to add information which will be used by the model to make correct predictions by estimating the conditional probability of a tag for a given sequence of words. The model uses a feature function that have multiple input values by modeling input sequence of words as features. The probabilities of assigned tags in output sequence are normalized over entire input sequence. This CRF based supervised machine learning algorithm is trained to find optimal weights of different features to generate an efficient algorithm that can be used to compute the most probable sequence of tags for new unknown input sequence. Stanford NER is available for various different languages. In this thesis, Stanford NER for English language has been used.

As mentioned on their website, Stanford NER provides three different models that cover different numbers of generic entities:

[1] https://nlp.stanford.edu/software/CRF-NER.shtml

3 class: Location, Person, Organization
4 class: Location, Person, Organization, Misc
7 class: Location, Person, Organization, Money, Percent, Date, Time

This thesis requires identification of location entities and as it is already covered by 3 class model so 3 class Stanford NER model is used in this thesis. This 3 class model is trained on the mixture of modern corpora such as CoNLL, MUC-6, MUC-7 and ACE and some limited amount of in-house data. Stanford NER can be trained on any kind of labeled data to train custom models as well. Stanford CoreNLP toolkit (version 3.9.1) is used in this thesis [36].

spaCy NER :

spaCy v2.0 is implemented in Python. spaCy NER [25] is a fast statistical model and every decision made by the model is prediction. However, the exact working of models is not described in any publication. However, in the website[2], the author mentions that, "spaCy NER applies sophisticated word embedding using subword features and 'Bloom' embeddings, a deep convolution neural network with residual connections, and a novel transition-based approach to name entity parsing." spaCy provides packages in multiple languages and English language has 4 packages. In this thesis, *'en_ core_ web_ sm'* is used for the purpose of entity extraction and spaCy NER's entity type 'LOC' is taken into consideration to identify locations.

NLTK NER :

NLTK[3](version 3.5) is a python based NLP toolkit [5]. NLTK NER uses Maximum Entropy classifier. This algorithm requires text to be tokenized, tagged with Part-of-Speech tags and then a parser chunks the tokens based on their Part-of-Speech tags to find named entities. This maximum entropy classifier is trained on ACE corpus. Apart from 'LOCATION' entity type, it also outputs 'GPE' entity type. To recognize location entities from the narrative text, outputs of above two entity types are combined.

AllenNLP NER :

AllenNLP[4] provides two different NER models: Elmo NER and fine-grained

[2]https://spacy.io/
[3]https://www.nltk.org/book/
[4]https://allennlp.org/

NER. Elmo NER implements semi-supervised sequence tagging with bidirectional language models [44]. It uses pretrained GloVe Vectors for token embedding and a Gated Recurrent Unit (GRU) character encoder as well as a GRU phrase decoder. It is trained on CoNLL-2003 NER dataset and it achieves state-of-the art results. Whereas fine-grained-NER reimplements Lample [32] and uses a bidirectional LSTM with a CRF layer which detects a 16 semantic types in the text. This model is trained in the Ontonotes 5.0 dataset and it has a dev set F1 of 88.2. In this thesis, AllenNLP Elmo-NER (version 2.2.0) is used. This model outputs 'GPE' entity type for the location entities.

The overview of various pretrained NER models used in thesis is shown as below:

Table 3.1: Overview: NER Model Performance

NER model	Data	Model	F1 score
NLTK	ACE 2004	MaxEnt classifier	0.89 ±0.11
Stanford NER	CoNLL, MUC-6 MUC-7 and ACE	CRFClassifier	87.94 %
spaCy	OntoNotes	Multi-task CNN	85.85 %
AllenNLP	CoNLL	ELMo	90.87 ±0.13

3.2 Lexical Resources

WordNet:

WordNet[5] is a large lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets) inspired by psycho-linguistic theories of human lexical memory. Each synonym set (synset) represents one underlying concept. WordNet interlinks words based on semantic relation among the words, that is, depending on the specific sense of words. In other words, words that are in the close proximity are semantically disambiguated. WordNet synsets are linked to each other by conceptual relations. One of the most frequently encoded relation between synsets is the super-subordinate relation (hyponymy). It forms a hierarchy for nouns which ultimately go up to the root node. Instances are always leaf nodes in their hierarchies. Verb synsets are also arranged in a hierarchical manner whereas

[5]https://wordnet.princeton.edu/

adjectives are organized in terms of antonymy. In this thesis, WordNet interface[6] from NLTK toolkit is used.

VerbNet:

VerbNet[7] is a largest network of English verbs that links the syntactic and semantic information of the verbs. VerbNet has explicit links to other lexical resources such as WordNet and FrameNet. VerbNet is inspired by Levin verb classes to systematically construct lexical entries where each VerbNet class is described by thematic roles, selection restrictions on the arguments and frames. VerbNet has 274 classes and 3769 lemmas. There are total of 23 thematic roles in VerbNet and each argument is assigned a unique thematic role. When verbs have more than one senses, they are assigned different thematic roles. Each class contains identifier such as 'Escape-51.1'. These classes contains members that has same semantic-syntax behavior such as class 'Escape-51.1' has associated 22 members such as ARRIVE, EMIGRATE etc and 5 thematic roles. In this thesis, VerbNet interface[8] from NLTK toolkit is used.

FrameNet:

The FrameNet project[9] is a lexical database of English that aims at implementing the thoery of Fillmore's Frame semantics. The syntactic environment of words are generally systematically aligned with the semantic frames that are evoked by words. A Frame is a conceptual structure describing a particular type of a situation or an event along with the required participants. A frame contains "frame elements" which are the roles of a frame and "lexical units" which are the frame evoking words. In this thesis, FrameNet interface from NLTK toolkit is used. There are several relations between frames which are included in FrameNet. The most important relations explained in official FrameNet interface documentation[10] are as below:

- Inheritance: Parent-child frame relation where child frame is a subtype of a parent frame.
- Using: The child frame presupposes the parent frame as background.
- Subframe: The child frame is a subevent of a complex event represented by the parent frame.

[6]https://www.nltk.org/howto/wordnet.html
[7]https://verbs.colorado.edu/verbnet/
[8]https://www.nltk.org/$_m$*odules/nltk/corpus/reader/verbnet.html*
[9]https://framenet.icsi.berkeley.edu/fndrupal/
[10]http://www.nltk.org/howto/framenet.html

SemLink

SemLink[11] links information provided by VerbNet and FrameNet and generates many-to-many mapping between these two lexical resources. It generates linking in two parts[12]:

1. **VerbNet Class / FrameNet Frame Mapping**
   This is a many-to-many mapping where one VerbNet member can map to more than one FrameNet frames and one FrameNet Frame can map to more than one VerbNet member. The attributes of this mapping are as below:

   Table 3.2: VerbNet Class / FrameNet Frame Mapping

Attribute Name	VN/FN element
class	VerbNet class ID
vnmember	VerbNet class member
fnframe	FrameNet Frame
fnlexent	FrameNet lexical entry ID
versionID	VerbNet version ID

   The structure can be demonstated by this example :

   <vncls versionID="vn1.5" fnlexent="" fnframe="DS" **vnmember="leave" class="13.3"**/>

2. **VerbNet ThematicRole / FrameNet FrameElement Mapping**
   This mapping includes the possible role correspondence between the VerbNet classes and frames of the FrameNet. The role in the FrameNet Frames are expressed as Frame Elements. The attributes of this mapping are as below:

[11] https://verbs.colorado.edu/semlink/
[12] https://verbs.colorado.edu/semlink/semlink1.1/vn-fn/README.TXT

Table 3.3: VerbNet ThematicRole/FrameNet FrameElement Mapping

Attribute Name	VN/FN element
fnrole	FrameNet Frame Element
vnrole	VerbNet Thematic Role

The structure can be demonstrated by this example :

```
<vncls fnframe="Departing" class="51.1">
   <roles>
      <role vnrole="Theme" fnrole="Theme"/>
      \<role vnrole="Location" fnrole="Source"/>
      <role vnrole="Location" fnrole="Path"/>
   </roles>
</vncls>
```

Chapter 4

Corpus

This chapter describes the historical dataset used in this thesis i.e., the narrative named *The Travels of Marco Polo* and its analysis in depth. Section 4.1 describes the steps taken for the preparing the data that can be further analyzed. Section 4.2 describes the additional domain knowledge based resource gazetteer, the need for it and the preparation of it. Section 4.3 describes the gold standard setup that will be used to evaluate the performance of the algorithm at various stages and finally, Section 4.4 describes the statistical analysis of the dataset along with peculiarities observed in the dataset.

The Travels of Marco Polo is a 12th century travelogue written down by Rustichello da Pisa from the stories told by Italian explorer *Marco Polo* who narrated his own travels through Asia between 1271 and 1295. The original narrative is written in Franco-Italian and translated in English by various authors including 18th century translation by Hugh Murray and Henry Yule and these both translations are used in this thesis.

Pages from both the books are available as a scanned resources as well as text resources but these text resources were erroneous and hence, scanned resources were transformed via Optical Character Recognition (OCR) technology into textual representation that can be searched and indexed for the further processing. In this thesis, two different English translations, by Hugh Murray and Henry Yule, are used for different purposes. It uses *The Travels of Marco Polo* by Hugh Murray as a main text resource for the identification of the place references visited by *Marco Polo* and also for the construction of gazetteer. The another translation by the author Henry Yule is used only as an additional resource for gazetteer construction. The book by Hugh Murray has been organized into three parts and these three parts are further divided into sections as shown in the Table 4.1.

Table 4.1: Overview of the narrative

PART	Description	Total Sections
PART I	It describes the travels in China	81
PART II	It describes the travels in Central Asia	51
PART III	It describes the voyage through Indian Seas and the journey back to Venice	61

4.1 Data Preparation

In this section, the process of automatic setup of data preparation and data cleaning has been described as below:

4.1.1 OCR

The scanned PDF is converted into searchable text for further processing using the technology of Optical Character Recognition. As a first step, manually all three parts of the books were separated and fed to the OCR tool. OCR tool first converted each page into corresponding image and then each image into corresponding simple text. The text from every pages are concatenated for each part of the book and as a result, three text files corresponding to each part of the book are generated.

4.1.2 Context Sensitive Spelling Correction

When the scanned PDF is converted into a text file, it contains several errors depending on the quality of a scanned PDF. These errors have to be fixed manually before the corresponding text is used for further processing. As shown in the below Figure 4.1, the OCR errors have occurred while converting scanned PDF into searchable text. This example also shows that whenever there is a page break mid-sentence, the header of the page is included as a part of the sentence which has to be removed manually.

In addition, the book contains several footnotes which provides additional information about certain text in the narrative. However, these footnotes are not useful for the task which needs to be solved and hence, these footnotes and their indicator (* sign or numbers as superscripts) need to be removed manually as well. For example, in the Figure 4.2, the author provides further information about the ship material *fir* in the footnote as shown in Figure 4.3 which is not useful at all for the task that is being solved.

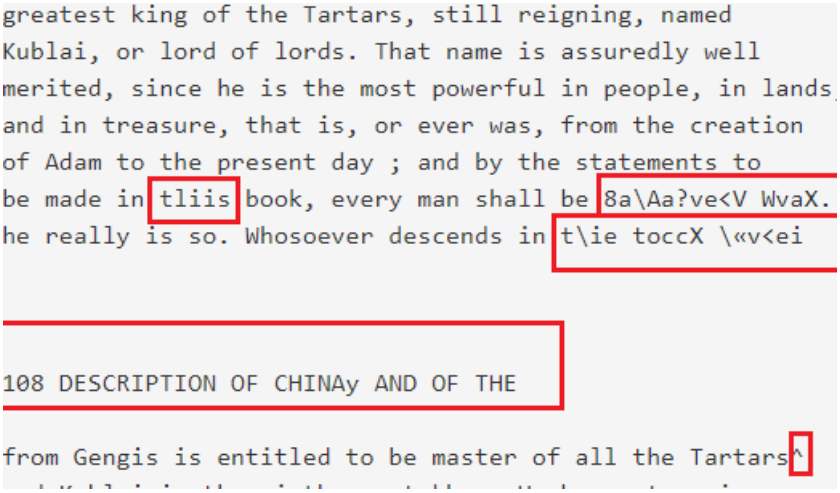

Figure 4.1: Context specific errors in OCR

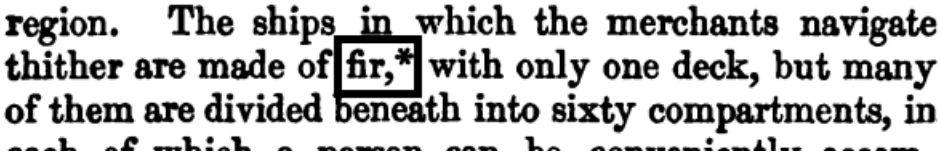

Figure 4.2: An example of footnote indicator

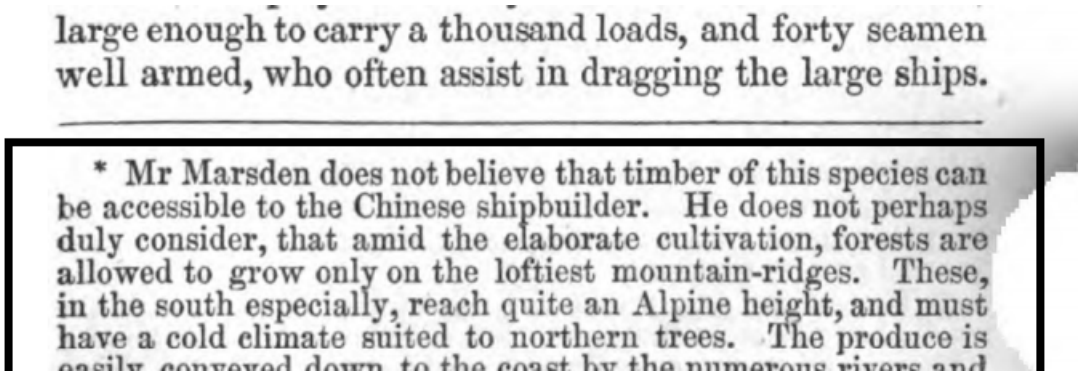

Figure 4.3: An example of an irrelevant footnote

4.1.3 Section Splitting

As described in the beginning of this chapter, this book is arranged into three parts corresponding to three major geographical divisions and these parts are further divided into sections where each section represents either the internal journey within the main geographical division described in Table 4.1 or it represents significant information or specific event corresponding to a certain location within that main geographical division. Hence, it is observed that section title is an important information to retain as it contains the information related to the places visited by *Marco Polo*. For that purpose, the text file corresponding to each part is further divided into several text files for each section where the name of text file represents the section number, the section name and, it contains the content of that particular section.

4.1.4 Sentence Splitting

From the above processing, content of the narrative is organized into several text files corresponding to sections while retaining important information such as section details along with manual contextual spelling corrections of the text. Now this text is further divided into a sentence level granularity where each sentence will be analyzed to find necessary information in the context of travel. For that purpose, a comma separated file is generated for each part of the book with the sentence level granularity and section details corresponding to each sentence. A pre-trained Punkt Sentence Tokenizer for English language from Natural Language Toolkit (NLTK) [1] is used for sentence splitting. These sentences were further manually examined to correct splitting mistakes.

4.1.5 POS Tagging

The sentences are preprocessed for the purpose of Part-of-Speech tagging which helps to extract information like entities, events etc. Stanford Part-of-Speech tagger [36] was used for this purpose. An example of the Part-of-Speech tagging outcome from Stanford tagger has been shown in the following example.

[PRP He] [VBZ resides] [IN in] [DT the] [JJ vast] [NN city] [IN of] [NNP Kambalu]

Here, each word is assigned its part-of-speech tag. A list of Part-of-Speech tags is given in Appendix A.

[1] https://www.nltk.org/api/nltk.tokenize.html

4.2 Gazetteer Preparation

Gazetteers are reference lists of entity names that are already labeled relevant to the task. The task of entity extraction can often benefit from gazetteers where entity extraction is very challenging due to scarcity of proper resources for particular domain specific knowledge and gazetteers encode additional background knowledge for a certain domain specific task.

In case of the historical text, there is an absence of a universal tooling, even for the text as late as 20th century. There is also a lack of a common standards for gazetteers. Hence, for the 12th century travelogue of *Marco Polo*, the index from the book as a knowledge base for the gazetteer construction has been proposed in this thesis. Two different translations of the travelogue by two authors, Hugh Murray and Henry Yule, with a significant difference in the length of index has been considered for gazetteer construction as shown in Table 4.2.

Table 4.2: Gazetteer generation

Gazetteer Type	Total Unique Entities
Index of Hugh Murray	289
Index of Henry Yule	3317
Combined Index of Murray and Yule	3498

Both indexes are alphabetical list of entities with corresponding page references. These entities can be names, places, events or any other terms related to the content of the book. Also, entities of the index could be single word or multi words, offering alternative names for the corresponding entity. The indexes are available as scanned PDF and they are converted to searchable text using the OCR tool described in Section 4.1.1 and also manually examined for context sensitive spelling correction as described in Section 4.1.2. The text file corresponding to entire index is then converted into multiple text files according to alphabetical order A- Z. An entity within index contains alternative names, multiple references along with page number as shown in below example:

ABASCIA (Abyssinia), kingdom of,324. The inhibitants converted by St Thomas,325. Its king defated the ruler of Adel (Aden), 326.Productions of the country, 327.Abraiamain (Bramins), order of, 293,304-308.

Here, entity name is *ABASCIA*. The alternative name for entity is given in parentheses pair ( ) as *Abyssinia* and it contains 5 references along with page numbers. First, all the entries in the index are organized as Entity Name,

Alternative Name, References and Page Numbers as shown in Table 4.3. Here, it is important to retain the entity and its alternative name but descriptions and page numbers are not needed in generated gazetteer for the task of an entity recognition so they will be discarded later while annotating entities.

Table 4.3: Organizing entries in Index

Entity	Alternative Name	References	Page No.
ABASCIA	Abyssinia	kingdom of	324
ABASCIA	Abyssinia	The inhibitants converted by St Thomas	325
ABASCIA	Abyssinia	Its king defated the ruler of Adel (Aden)	326
ABASCIA	Abyssinia	Productions of the country	327
ABASCIA	Abyssinia	Abraiamain (Bramins), order of	293,304-308

4.2.1 Ambiguity Resolution

As discussed above, three versions of gazetteers are created out of indexes of two different translations. Historical translations are often erroneous and contains several ambiguities which is described in detail in Section 4.4.2. And hence, when two different indexes of different length are combined, it requires manual efforts to solve these ambiguities which is done by the author of this thesis. Here is the one example of a character "Alau" as shown in Table 4.4 and Table 4.5 for Murray and Yule translations respectively. This character has some variations as shown in Alternative Names but all these variation refer to the same person "Alau" and it looks like as shown in Table 4.6 after ambiguity resolution.

Table 4.4: Generated gazetteer from Murray translation

Entity Name	Alternative Names	Entity Tag
Alau	Hookalu	Person

Table 4.5: Generated gazetteer from Yule translation

Entity Name	Alternative Names	Entity Tag
Hukalu Khan	Alau,Hukalu	Person

Table 4.6: Gazetteer after ambiguity resolution

Entity Name	Alternative Names	Entity Tag
Alau	Hookalu,Hukalu,Hukalu Khan	Person

4.3 Gold Standard Corpora

Trustworthy corpora is very important for the meaningful evaluation of algorithms. These collection of standard and trustworthy annotations are called gold standard corpora and the quality of human annotations as a gold standard directly affects the evaluation of algorithms. In this thesis, several manual annotations are established by author manually as gold standards which are used further for the analysis of various NLP techniques and algorithms at various steps.

First established gold standard is for the named entity, specifically, place references. Each place reference is annotated with its granularity level and boundaries. For named entities, it is important to define boundaries of annotations clearly and hence, each annotation clearly defines the boundary of a place reference. By manually examining the named entities for this thesis, it is found that majority of the named entities contain one or two tokens and very few entities have three tokens and there is almost no entity with a token length of more than three so maximum number of chosen tokens for named entities in this thesis is three.

In addition, motion verbs work as a motion event triggers and hence, motion verbs are manually annotated in this thesis which will be used further for the evaluation of the motion verbs as motion event triggers. For each part of the book, count of motion event triggers are as shown in Table 4.7. Table 4.7 shows that there are 69 events related to motion in Part 1, 53 events related to motion for Part 2 and 93 events related to motion in Part 3.

Table 4.7: Count for Motion Event Triggers

PART	**Motion event triggers count**
PART 1	69
PART 2	53
PART 3	93

Lastly, the goal of this thesis is to find the visited locations by *Marco Polo* and hence, it is important to manually set a gold standard for the travel path to evaluate the retraced travel path. For the purpose of that, locations traveled by *Marco Polo* were found and connected in a chronological order. Furthermore, this gold standard annotated travel path is also mapped manually on a contemporary world map to confirm it with the work of other researchers found from various sources on the Internet as shown in below Figure 4.4. There is a different of opinions among various researchers regarding the path taken by *Marco Polo* and hence, the author of the thesis had to rely on the various sources on the Internet.

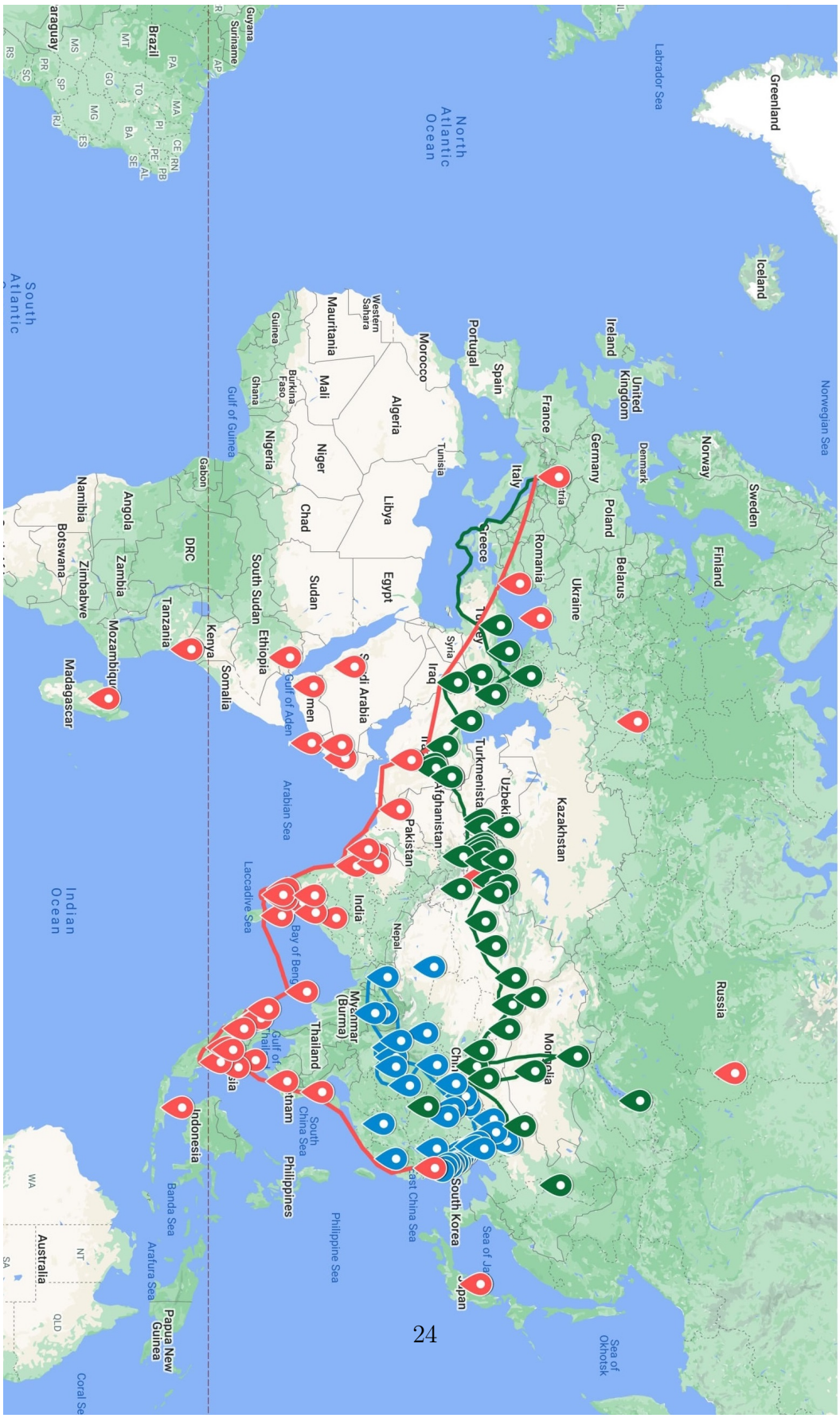

Figure 4.4: Gold standard annotated travel path

4.4 Data Analysis

Quantitative and qualitative data analysis play an important role in understanding of the narrative text. Hence, this section describes the various statistical and qualitative analysis performed on the narrative text.

4.4.1 Quantitative Data Analysis

Corpus statistics of an unannotated narrative corpora is important to understand domain-specific terminology and also to understand the size of the corpora. This thesis analyzes the descriptive information about the book to get insights about the structure of the book as shown in Table 4.8. The summary of the unstructured text provides a big picture of the data.

Table 4.8: Descriptive information about corpus

PART	Total sections	Total sentences	Average no of sentences	Total words	Total unique words
PART 1	81	971	11.9877	26140	3663
PART 2	51	592	11.6078	14563	2466
PART 3	61	861	14.1148	20220	3228
Total	193	2424	12.5596	60923	5793

Furthermore, in NLP, a corpus is made of sentences and sentences are collections of tokens. Analysis of these tokens as an n-grams, specially $n \geq 1$ are short phrases that give us information about central themes in the narrative. For example, most common bi-gram is "great khan" which denotes that *Kublai Khan*, who is also referred as "great khan", is one of the central character in the narrative. Here, most common uni-gram, bi-gram and tri-grams are analyzed respectively as shown in below Figures 4.5, 4.6, 4.7.

Figure 4.8 shows the distribution of Part-of-Speech tags in entire narrative. This is very useful because the task of entity extractions deals with NN (noun) tags and the task of events extractions deals with various tags related to verbs.

Also, the goal here is to find the locations traveled by *Marco Polo* and hence, the motion event triggers play an important role in identifying traveled locations. Figure 4.9 shows the coverage by two different lexical resources, VerbNet and FrameNet, for the motion verbs as motion event triggers. There are total of 59 unique verbs that represent the motion event. 54 of them can be identified by both lexical resources, 2 motions verbs are covered only by VerbNet and 3 motion verbs are only covered by FrameNet. Figure 4.10 describes the most common verbs in narrative based on the frequency of occurrence.

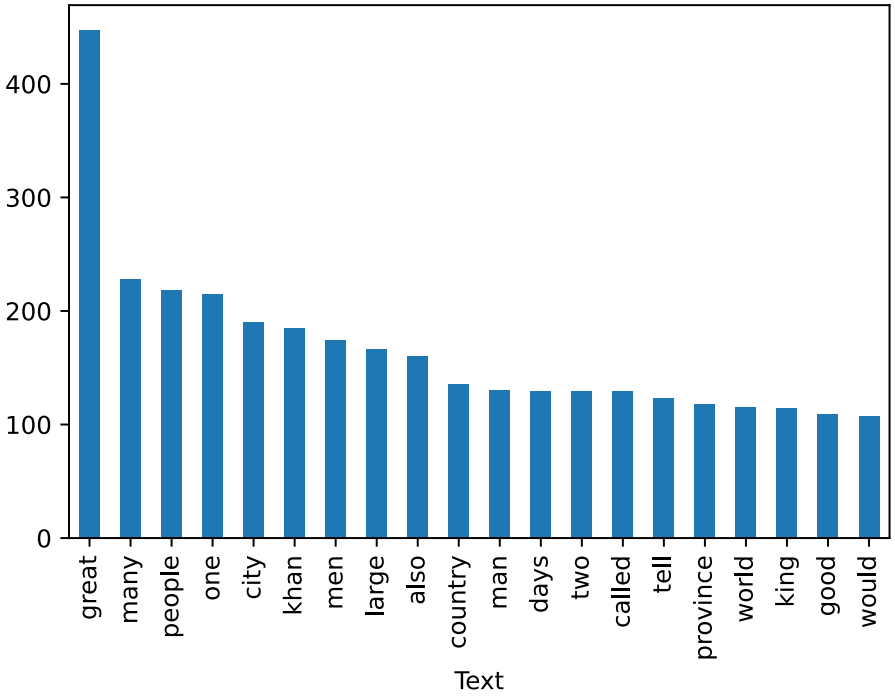

Figure 4.5: Most common uni-gram

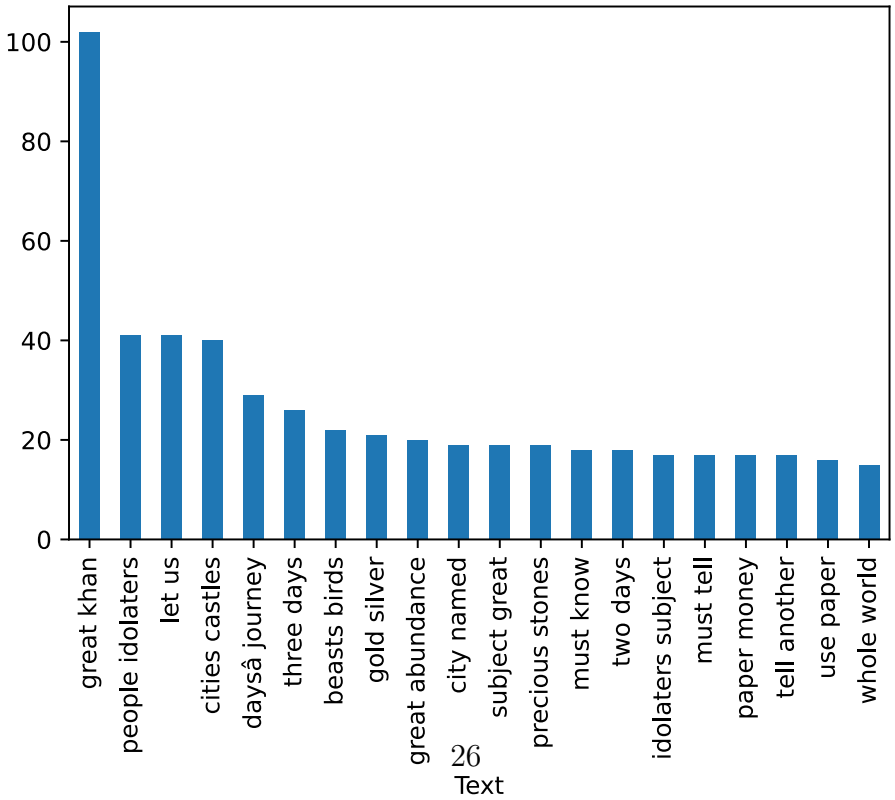

Figure 4.6: Most common bi-gram

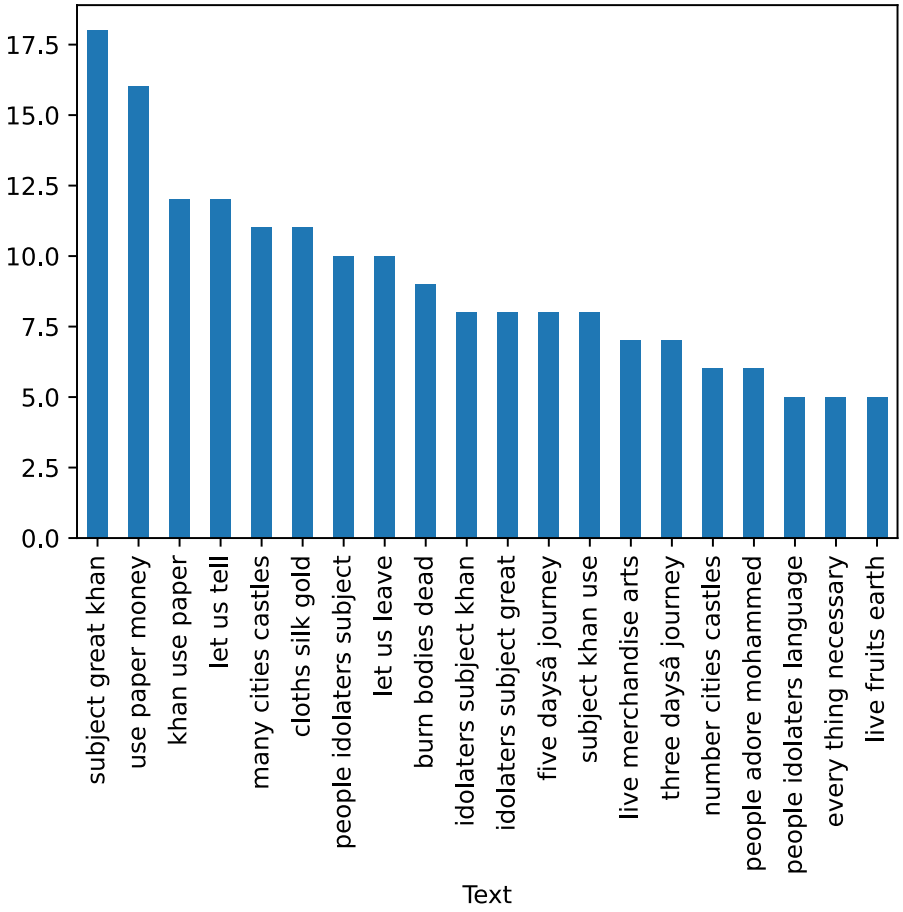

Figure 4.7: Most common tri-gram

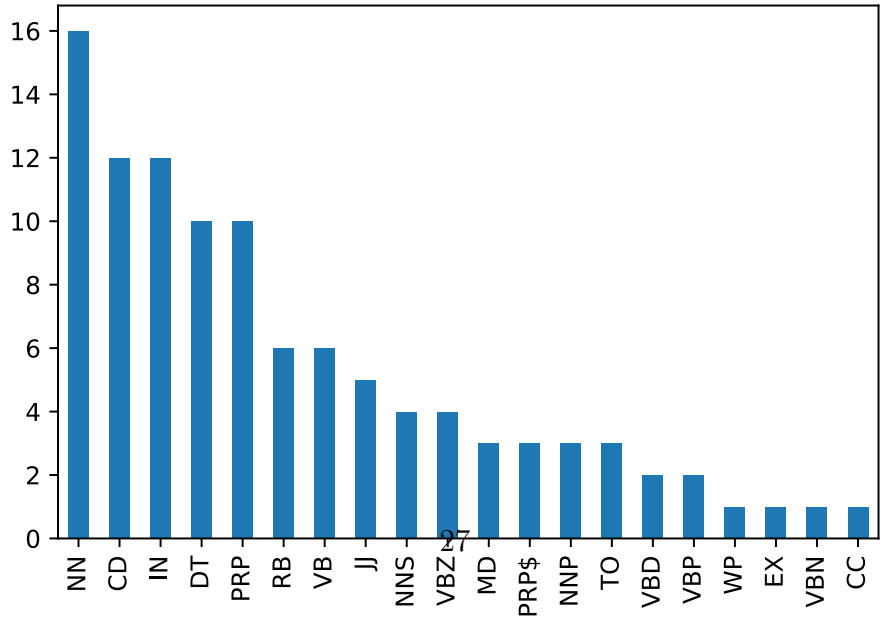

Figure 4.8: POS tags distribution

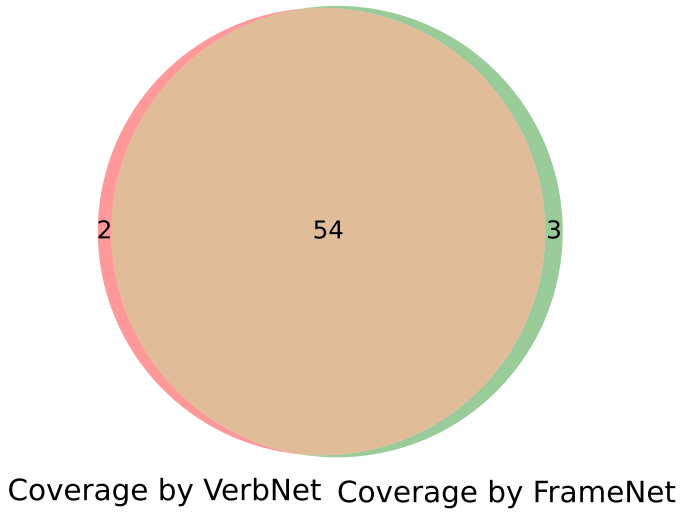

Figure 4.9: Motion verb converge by VerbNet and FrameNet

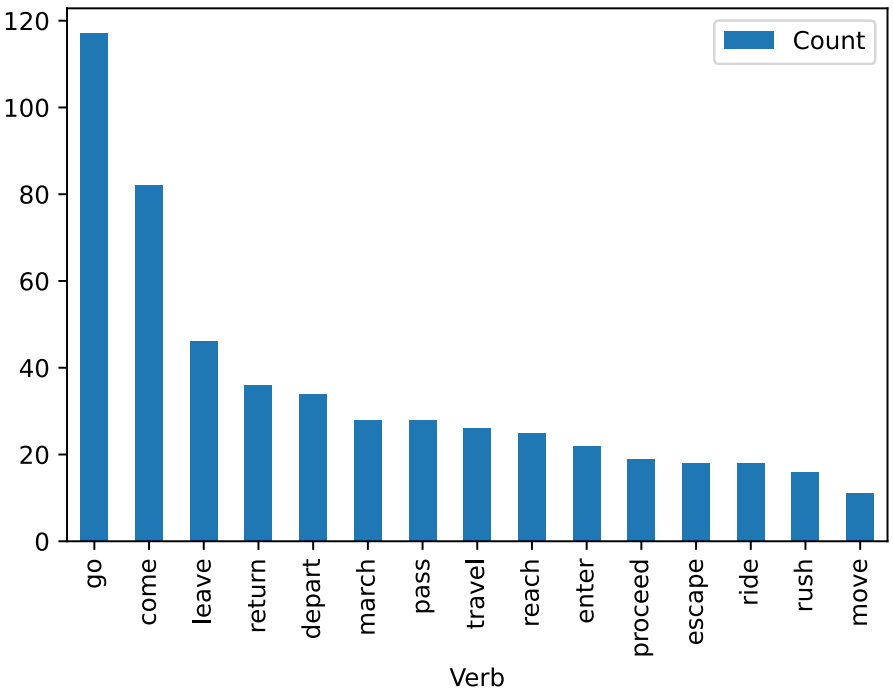

Figure 4.10: Most common motion verbs

4.4.2 Qualitative Data Analysis

This section describes the various challenges faced with historical dataset as well as observed domain-specific dataset phenomena that ultimately affects the quality of the results.

Challenges With Historical Dataset

Historical languages are quite different from its modern equivalent in several linguistic aspects such as syntax, lexicon, semantics and morphology, and hence it presents significant challenges for modern NLP tools developed for contemporary languages. In addition, OCR quality is generally quite poor and presents a number of problems, especially with translation of historical texts. As Figure 4.1 shows, OCR systems [21] have problems recognizing non-standard characters. In addition, OCR systems are also unable to separate marginal notes from the main body of the text which makes incoherent sequence of words and possibly breaks the continuity of the sentence or the paragraph.

There are certain challenges present with the translations of historical texts such as individual interpretation of the writer and capitalization. Errors in capitalization present significant challenges for the machine learning systems because in modern English, all nouns are capitalized and algorithms have learned that so when nouns are not capitalized, it creates problem for NER tools. Another issue arises from the lack of standardization process for historical texts which results in spelling variation based on the individual preference of the writer and the dialectal influences. This problem is clearly visible when two different translations of *The Travels of Marco Polo* are compared. As the name suggests, additional variance introduced by spelling variation has further consequences on almost all the processing of the historical data. If all the variants of an entity are treated as a different identity, then it results into missing out on useful information. In addition, knowing all the variants of an entity and individually specifying all the variants is highly impractical. If the case of Part-of-Speech tagging is considered: if a given word occurs with 10 different spelling variants, the required amount of training data increases by tenfold and this creates a severe problem when there is a sparse amount of training data. Similarly, lack of standardization in the use of punctuation marks and the inconsistent use of spacing can introduce additional variance and ambiguity with regards to sentence boundaries. However, problematic punctuation and spacing occurs much less frequently than spelling variation and it is easy to standardize them manually. In addition, naming convention keeps changing over time,e.g., this historical corpora contains special characters like ' - ' as part of the various entities such as *Kain-du, sa-fun-du.* Different naming

conventions in historical times makes it harder for modern English taggers to perform well on it, and in turn it affects the performance of modern NER tools.

Observed Dataset Challenges

These are the observations of the author of the thesis about the narrative of *Marco Polo.* Opinions of other researchers given here are taken from the various sources on the Internet[2] and it could be personal opinion of the authors in those sources.

- **Non-existing geographical entities:** Geography of locations gradually changes over time and hence, associated geographical data also change over time such as boundaries of geographical entities. In addition, formation of a new geographical entity and disappearance of an existing entity happens over time according to their past political, ethnic, linguistic divisions or certain unexpected natural calamities. However, they present a significant challenge when mapping them to their contemporary equivalents. Here, it is observed that there are many historical geographical entities that are difficult to locate either because they are non existent or their importance in a narrative is fairly small and hence, they are presented in a narrative without detailed descriptions but they are important for this thesis for retracing the travel path. For example, *"Greater India"* and *"Lesser India"* are two different countries in the narrative and these countries do not exist anymore. The geographical region presented by these countries are known by a different name in today's era.

- **Geographical renaming:** It was one of the challenging problem encountered in this thesis. Almost all the geographical references in the narrative have been renamed over time and it is difficult to establish a link between historical reference and their equivalent contemporary reference. The author couldn't find any previous research work that establishes links between them and relied on various sources on the Internet for establishing these links. For example, one source on the Internet has noted that "Location of '*BARSCOL*' is very unclear but is thought to be around the eastern end of the present day Tian Shan Mountains."

- **Change of boundaries:** It was another major issue encountered in this thesis. There were quite a few examples that described the geographical references that are spread into more than country. For example,one

[2]https://idlethink.wordpress.com/2008/08/31/indulgence-sin/

source on the Internet noted that, "Khorasan is a Pahlavi and Avestan word which means 'the land of sunrise'. *Greater Khorasan* included territories that presently are part of *Iran, Afghanistan, Tajikistan, Turkmenistan* and *Uzbekistan*".

- **Identifying contemporary equivalent of certain geograhical entities:** There are large number of instances of geographical references in the narrative that are difficult to locate in contemporary map. For example, certain locations were possible to map, like *"the city of Pein"* to the modern day closest equivalent of "*Yutianxian*". However, certain geographical references where much harder to connect with contemporary equivalent. One such example is *"the Province of Acbalec Manzi"*. One of the expert had noted that "As far as I can ascertain from the Yule-Cordier notes, it is the flat plains on the southwest march between Hanzhong and Chengdu."

- **Unnamed and descriptive place references:** There are mentions of certain places such as plains, mountain ranges, lakes or particular areas without any names. However, it is important to locate them for the reconstruction of the travel path. *"Plain of Bargu*" is one of the example which does not refer to any particular location and hence, it requires an input from an expert to decide the area from the description given in the text. One of the expert has made the note that "*Plain of Bargu* appears to be the same as that mentioned in Mongol history as Barguchin Togrum or Barguti, commencing about Lake Baikal, the river Barguzin that feeds it, and a town on its banks, Barguzinsk." Here, *Marco Polo* talks about an area which is somewhere around the location *Bargu* but exact size and location of that area is unknown.

- **Incorrect mapping of a border-line areas on a map:** Problems with large areas such as rivers and mountains is, there can be multiple countries or cities alongside them. So while mapping them on a google map, it can lead in a opposite direction with a large margin and hence, even though it is accurate, it leads to an error if tried to place them on a contemporary map.

- **Lack of specific details about vistied places:** The same issue also occurs with mention of the certain countries where *Marco Polo* has just visited border areas. However, mapping them on a map leads to false placing somewhere in the center of the country. One such example is the *"Province of Tebet"*. One of the researchers has noted that "Marco Polo is vague about the exact boundaries of what he experienced as Tibet;

Yule & Cordier suggest the boundaries were farther east than present day political borders, and also that Polo himself only saw a small part of it."

- **Mistakes of Marco Polo:** There are also significant non existing historical locations for which it is impossible to determine their identity. The reason could be that, as noted by some experts that mistakes or misjudgements of *Marco Polo* in recalling locations or describing the visited locations. Another possible reason could be the change of the language over time and the translations are highly dependent on the author's interpretation as well.

Chapter 5

Methodology

The following chapter presents my approach for the identification of the travel path of *Marco Polo* from his narrative text. For the objective of identifying the locations that were actually visited by *Marco Polo*, various location entities were identified using various Named Entity Recognition techniques. Also, to separate the stories related to the journey of *Marco Polo* from the other various stories in the text, motion event triggers were identified using several lexical resources and the best performing out of them was chosen. Then the relationship between those extracted location entities and motion event triggers was examined to find the locations that *Marco Polo* actually visited. Once the relationship is identified, the location entities, which are linked to motion events, are connected in a chronological order to generate a travel path.

A travel path of *Marco Polo* can be defined as a series of the continuous and consecutive locations visited by *Marco Polo* to reach from the starting location (Venice) of his journey to his destination location (China) and to return from China to Venice. Now several Natural Language Processing techniques discussed in the previous chapters have been used for the goal of travel path retracing. The requirements for detecting travel path are:

1. *Identifying locations entities:* Location entities are identified using the task of Named Entity Recognition which will be discussed in Section 5.1.

2. *Motion events detection:* Motion event denotes the action of travel from one location to another and the motion events are identified using motion event triggers which will be discussed in Section 5.2. In addition, the context in which event triggers are used is examined to identify the context of travel.

3. *Location entities and motion event linking:* After finding events and entities in the text, shallow parsing technique is used to identify the

relationship between the motion event triggers and the extracted location entities as discussed in Section 5.3.

5.1 Identification of Location Entities

For identification of a travel path, first of all, it is important to identify place mentions, whether they are visited by Marco Polo or they are part of the other stories and this has been implemented as the first step in this pipeline. Several Named Entity Recognition techniques have been implemented and evaluated for the identification of location entities in the unstructured text and the NER tools which performs the best will be selected for the further steps.

Named Entity Recognition is the task of identifying various entities in the text as discussed in Section 2.1 but the point of interest here is the location entities so this thesis will focus on location entities only. Using the manually annotated data as described in Section 4.3, the performance of state-of-the-art Named Entity Recognition tools are evaluated for the identification of location entities. This thesis chooses various pre-trained Named Entity Recognition tools that are state-of-the-art tools and they are representative of variety of approaches. It helps to know which approach works well for the 12th century old historical text. Some tools are based on supervised machine learning algorithm that considers linear sequence prediction models and leverages the extensive feature engineering or neural network models leveraging word embeddings whereas some tools are based on rule-based approaches as described in Section 3.1. Multiple Named Entity Recognition systems are considered based on below criteria: a) Systems that shown promising performance on standard corporas b) Systems that are widely and commonly used in NLP c) Systems that can be applied on any corpora, specially historical corpora d) Systems that are representative of different approaches.

These machine learning based Named Entity Recognition tools have achieved noteworthy results over modern day English corpora such as CoNLL with F1 score of 90.94 for recognizing various entities [32]. But as discussed in the previous Section 4.4.2, the English language has changed significantly over time and historical language was quite different than modern day language. For example, entity names never contain special characters in contemporary English whereas the translation of *The Travels of Marco Polo* has frequently used location names with special characters. This inconsistent naming conventions for entities is shown in the Example 5.1.

(Example 5.1) *"Five days' journey from Cian-glu is Cian-gli, where are many cities and castles."*

The following four pre-trained Named Entity Recognition tools: NLTK NER, spaCy NER, Stanford NER and allenNLP NER as described in Section 3.1 and gazetteer as an external resource as discussed in Section 4.2 have been used in this research.

- **Application of pre-trained NER Models**
  In case of the pre-trained Named Entity Recognition tools, it is possible to add new features on top of the standard features distribution and custom models can be trained on labeled data for the domain specific applications which can help to make the model more robust and independent. However, because of lack of sufficient training data for the historical text, specially related to the 12th centuty old narrative of *The Travels of Marco Polo*, it is not possible to train the custom model and thus, this research uses the provided pre-trained models for Named Entity Recognition. Pre-trained Named Entity Recognition models generally keep entity types generic. However, it is not an issue for this research as the entity of interest for this research i.e., location entity, is covered in generic entities provided by pre-trained models.

- **Gazetteer for the Named Entity Recognition**
  A Gazetteer is a reference list of entity names that are already labeled relevant to the task as discussed in Section 4.2 and it is an important knowledge resource for entity extraction. An index of a book is an alphabetical listing of various mentions and references such as names, places and things along with associated page numbers for each reference. This knowledge from index can be used to create name lists, known as gazetteer, that can be further used for the identification of entities. As discussed in previous Section 4.2 for gazetteer generation, three different versions of gazetteers have been generated and applied for the task of a Named Entity Recognition:

  1. Gazetteer generated from the index of book by Hugh Murray
  2. Gazetteer generated from the index of book by Henry Yule
  3. Gazetteer generated from the combined index of Murray and Yule

  The purpose of creating these three versions of the gazetteer is to analyze how difference in the size of the indexes in different book affects the results of gazetteers in terms of entity extraction. The below mentioned procedure is applied to all three different versions of the gazetteer to identify location entities.

Procedure :

1. Scan the input sentence (S) from left to right
2. Set N = 3
3. Extract the n-grams (G) from the sentence
4. For an instance (I) of n-grams (G), search instance in the *Entity Name* column in the generated gazetteer. If not available, find it in the *Alternative Entity Name* column in the generated gazetteer
5. If found, find the corresponding *Entity Tag*
6. If *Entity Tag* is location, mark instance (I) as a location entity
7. Remove the instance (I) from n-gram (G)
   G = G - I
   Go to step 4
8. Else set N = N - 1
   Go to step 3

In this procedure, N denotes the upper value on number of grams that will be identified starting from N = 1. So N = 3 identifies uni-gram, bi-gram and tri-gram entities. There have been almost none to rare entities with more than three tokens so here it identifies entities up to three tokens.

5.2 Identification of Motion Events

Event extraction is a task in information extraction where mentions of events are extracted from the text. In this thesis, the point of interest is the identification of motion events and using motion events to further identify associated arguments to it. Now each event is identified by an event indicating word which is known as a trigger. This trigger could be tensed or un-tensed verbs, adjectives, predictive clauses or predictive phrases. So here, an un-tensed verb will be used to identify motion events using approaches based on various lexical resources such as VerbNet, FrameNet and WordNet. Let us consider a sentence shown in Example 5.2, which has three motion events triggers, namely 'leaving', 'sailing' and 'reach'.

(Example 5.2) *"On leaving that port, and sailing west and somewhat southwest 1500 miles, you reach a country named Cianba."*

Now to identify un-tensed verbs related to motion, first, all the verbs are identified in a sentence using Stanford Part-of-Speech Tagger[1]. To implement Stanford Part-of-Speech Tagger in python, python interface[2] Stanford CoreNLP is used. Identified verbs are then converted to their base form using WordNet Lemmatizer interface provided by NLTK[3] and then each of these verbs is classified into the category of motion verb or non-motion verb. The base form of the verb is root of a verb without any endings like s, ed, or ing. These verbs with various ending represent verbs in various tenses and Part-of-Speech tags related to verbs that were taken into consideration are as shown in Table 5.1.

Table 5.1: Part-of-Speech tags for verbs

POS Tag	POS Tag Description
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present

Stanford POS Tagger identifies POS tags for all the tokens and tokens with tags listed in Table 5.1 are then converted into verb base form VB using Wordnet Lemmatizer. This process is shown for example sentence in Example 5.3.

(Example 5.3)
Original Sentence: *"After climbing down a mountain, they descended on a plain area."*
Processed Sentence:"*After climb down a mountain, they descend on a plain area."*

The need to convert verbs into their base form is that, they further needs to be analyzed using approaches based on lexical resources VerbNet, WordNet and FrameNet for the identification of motion events.

[1]https://stanfordnlp.github.io/CoreNLP/pos.html
[2]https://pypi.org/project/stanford-corenlp/
[3]https://www.nltk.org/*modules/nltk/stem/wordnet.html*

- **Use of WordNet for Motion Events Extraction**
  As discussed in Section 3.2, WordNet has synonym sets (known as *synsets*) that represents a group of synonyms. For the purpose of motion events identification, three seed words, namely, DEPART, TRAVEL, ARRIVE are chosen after performing the experiments with different travel related words as seed words. Figure 4.10 shows the most common verb in this narrative are "Go" and "Come" but overall inaccurate predictions increases with these words so these words were not chosen as seed words even though they are most common in the text. "Depart" and its synonyms can indicate the source location, "Arrive" and its synonyms can indicate the destination location and "Travel" and its synonyms represent motion in general.

  To check if the given target verb is a motion event trigger or not, a synset of that verb is generated. Now semantic similarity between this synset of a target verb and synset of each of the seed word is computed. This semantic similarity between two synsets S1 and S2, is calculated using measure suggested by Wu and Palmer [60] which is based on the path length. It takes into consideration the number of nodes in each synset and the shortest path between them as

  $$\text{Sim(S1,S2)} = 2 \times \frac{N_3}{N_1+N_2+2\times N_3}$$

  N1 is the length of path between S1 and S3, N2 is the length of path from S2 to S3, where S3 is their most specific common super class. N3 shows the global depth of the hierarchy and is the length of path from S3 to root of the ontology [55]. Now Wu-Palmer similarity between two synsets S1 and S2 is

  $$\text{Dist(S1,S2)} = 1 - \text{Sim(S1,S2)}$$

  WuPalmer similarity is calculated using the methods given in WordNet interface[4] provided by NLTK in this thesis. The similarity score can be $0 < score \leq 1$. The threshold here chosen as 0.5 after doing experiments and comparing results to find the ideal value of the threshold. If the semantic similarity score of a synset of an any of the seed word and a synset of a target verb is greater than threshold, then the target verb will be marked as a motion verb.

[4]http://www.nltk.org/howto/wordnet.html

This process can be described as an algorithm as follow:

Algorithm 1: Identifying motion verbs using WordNet

Input: A verb-list(V) for an input sentence(S)
Output: A list containing motion-verbs
$maxscore \leftarrow 0$
$threshold \leftarrow 0.5$
$seed - list \leftarrow DEPART, TRAVEL, ARRIVE$
/* Generate a list of synonyms for all the seeds */
for $seed \in seed - list$ **do**
  $seedsynset \leftarrow seedsynset \bigcup synset(seed)$**end**
  /* Generate a set of synonyms for each verb */
  for $verb \in verb - list$ **do**
    $verb - synset \leftarrow synset(verb)$
    /* Calculate a similarity between each pair of a synonyms set of a verb and seed */
    for $seed - item \in seed - synset$ **do**
      for $verb - item \in verb - synset$ **do**
        $score \leftarrow$
          $wu - palmer - similarity(seed\_item, verb\_item)$
          /* Find the pair with maximum score */
        if $score > max - score$ **then**
          $max - score \leftarrow score$
        end
      end
    end
  end
  /* Add verb to the list of motion verb if there is similarity above defined threshold */
  if $max - score > threshold$ **then**
    $motion - verb \leftarrow verb$
  end

- **Use of VerbNet for Motion Events Extraction**
  As discussed in Section 3.2, the lexical resource VerbNet divides verbs into classes based on their syntax-semantics linking behavior. Each class contains identifier such as 'Escape-51.1'. These classes contains members that has same semantic-syntax behavior such as class 'Escape-51.1' has associated 22 members such as ARRIVE, EMIGRATE etc. Class 51 has further 10 sub classes that were identified as an indicators for the motion events and hence, class 51 is chosen as a seed for VerbNet. Now for each

target verb, all possible VerbNet classes associated with that target verb are identified and if any of the associated class belongs to the category of class 51, then target verb is marked as a motion verb. VerbNet classes associated with a verb can be identified using methods given in VerbNet interface[5] provided by NLTK.

- **Use of FrameNet for Motion Events Extraction:**
  As discussed in Section 3.2, FrameNet is a semantic dictionary of English, based on the concept of semantic frames. Each frame comes with a set of semantic roles describing the participants and objects of that situation. For the motion event identification, five frame elements have been assumed as seeds:

  – SOURCE: "The source is a location from where theme changes its location"

  – GOAL: "The goal is the location where theme ends up at the result of location change"

  – PATH: "Path is a ground over which theme travels from source to destination"

  – PLACE: "Place is a location without a specific path for theme"

  – DIRECTION: "Direction is a straight line between source location and destination location"

  Here, theme is an entity that is not a self-mover but does the motion and changes the direction. Now using these seed elements, a list of frames is generated which contains at least one of the seed element as a frame element and this list of frames is marked as a "motion event frames list". To classify a target verb, all frames associated with a target verb are identified. If any of the associated frame is in the "motion event frames list", then the target verb is marked as a motion event trigger. This method is implemented using various methods given in FrameNet interface[6] which is provided by NLTK.

Now when verbs are taken into consideration, it is very important to note that, the verbs can have multiple meanings depending on the various contexts in which they are being described and hence, it is difficult to map them to preexisting categories without the reference of the context in which they are being used. Here is one example from the narrative of *Marco Polo*. The FrameNet identifies the verb "descend" as the motion event trigger. Four

[5]https://www.nltk.org/*modules/nltk/corpus/reader/verbnet.html*
[6]http://www.nltk.org/howto/framenet.html

different frames contain this verb "descend" as a lexical unit as shown in below Table 5.2:

Table 5.2: FrameNet Frames for example verb

Lexical Unit	**Frame**
descend (on).v	Arriving
descend.v	Path_shape
descend.v	Traversing
descend.v	Motion_directional
descendant.n	Kinship

Frames *Arriving* and *Path_shape* represent the motion event in which an object theme moves in the directional of a goal. Whereas frame *Motion_directional* represents that object theme moves in a certain direction often by other forces and object theme is not necessarily a selfmover. The difference between these two can be explained by the below examples. Example 5.4 shows that object theme moved towards some goal (location) and Example 5.5 shows the direction of natural lineage.

(Example 5.4) *"After climbing down a mountain, they descended on a plain area."*

(Example 5.5) *"In this province there is a king named George, descended from that prince, and who indeed enjoys his power."*

So these lexical resources without any context identify all possible motion event triggers but they can be classified as a motion event trigger only after examining the context in which they are used. So now, instead of identifying the specific context for each verb, SemLink Project, described in Section 3.2, will be used to for classifying context of possible motion event triggers into travel context or non-travel context. SemLink links information provided by VerbNet and FrameNet and generates many-to-many mapping between these two lexical resources in two parts and context is identified using these mappings:

1. **VerbNet Class / FrameNet Frame Mapping**
   This is a many-to-many mapping between Verbnet and FramNet and using this mapping, all possible pairs of (VerbNet class, FrameNet frame) are identified for a possible motion event trigger (target verb). Now consider the example sentence given in Example 5.6.

(Example 5.6) *"He left from the country of Cianba."*

In this example, we have just one verb 'left' and a base form of this verb is 'leave'. Based on VerbNet Class / FrameNet Frames mapping, all possible linking between these them is as below:

<vncls **class="13.3"** vnmember="leave" **fnframe="DS"** fnlexent="" versionID="vn1.5"/>
<vncls **class="15.2"** vnmember="leave" **fnframe="DS"** fnlexent="" versionID="vn1.5"/>
<vncls **class="51.1-1"** vnmember="leave" **fnframe="Departing"** fnlexent="8622" versionID="vn1.5"/>
<vncls **class="51.1-1"** vnmember="leave" **fnframe="Path_shape"** fnlexent="1064" versionID="vn1.5"/>
<vncls **class="51.2-1"** vnmember="leave" **fnframe="Departing"** fnlexent="8622" versionID="vn1.5"/>

So here, for VerbNet member 'leave', we have five different results. Out of these results, five (VerbNet class, FrameNet frames) are found as highlighted in above results which will be used in the second mapping of VerbNet thematic roles and FrameNet frame elements for further processing.

2. **VerbNet ThematicRole / FrameNet FrameElement Mapping**
Using this mapping, all pairs of (VerbNet thematic roles, FrameNet frame elements) are identified for each of the (VerbNet class, FrameNet frames) pair obtained from previous mapping. This mapping generates the possible role correspondence between the VerbNet classes and frames of the FrameNet.

   In VerbNet, for a given verb class, thematic roles to each syntactic argument is assigned. There are 23 Thematic Roles in VerbNet and only 3 roles out of these 23 roles are of interest here that denotes the context of motion. These three roles are shown as below in Table 5.3 with examples.

   In FrameNet, these semantic roles are described as a Frame Elements. There are many different types of frame elements but here, we are interested in three types of Frame Element related to motion and locative entity as shown below in Table 5.4 with examples.

Table 5.3: VerbNet thematic roles for locative entity

Thematic Role	**Example**
Source	He DEPARTED **from Cianba**
Destination	He ENTERED **India**
Location	He spent one month **in the country of Cianaba**

Table 5.4: FrameNet frame elements for motion

FrameNet frame elements	Example
Source	He DEPARTED from the city of **Cian-fu**
Goal	He REACHED at the border of **India**
Undesirable_location	He ESCAPED from **China**

Now if Example 5.6 is continued, then the five pairs (vnclass, fnframe) are obtained from VerbNet class / FrameNet mapping. Now each (vnclass, fnframe) pair can be used as an input in VerbNet ThematicRole / FrameNet FrameElement mapping to extract pairs of (vnrole, fnrole). So here, for Example 5.6, several pairs of (vnrole, fnrole) are found. But only one pair of (vnrole, fnrole) is found where VerbNet thematic role and FrameNet frame element belong to the motion event related roles described in Table 5.4 and Table 5.5 respectively. The structure of this one pair is as shown below:

```
<vncls class="51.2" fnframe="Departing">
    <roles>
        <role fnrole="Theme" vnrole="Theme"/>
        \<role fnrole="Source" vnrole="Source"/>
</roles>
</vncls>
```

Now from the above listing, the context of the possible motion event trigger "leave" in Example 5.6 is determined. Here, VerbNet class is 51.2, FrameNet frame is "Departing" and, vnrole and fnrole both are of type "source". So this information indicates that this event trigger describes the event of motion and to be more precise, the event of departing from some location and that associated location entity with this verb is a source location from which theme (subject of an action) changes its location.

5.3 Location Entities and Motion Events Linking

From the previous steps, location entities and motion event triggers are obtained from the text. Now, a relationship between them is examined to check whether they are directly or indirectly linked. The requirement for this linking can be described via below Example 5.7:

(Example 5.7) *"Now I will leave this topic here and go to further describe the country of Cianba."*

Here, location entity is "Cianba" and motion event trigger is the verb "leave" but they are not directly or indirectly linked to each other and hence, it can be understood that, the mention of "Cianba" here is not related to the travel and it is part of the other stories in the narrative. In addition, sometimes the location entity is embedded in a noun phrase such as '*the country of Cianba*'. It is useful to extract such noun phrases as they do not only contain only location entity but they also embed additional information about the locative entity. For example, "*the country of Cianba*", expresses that location entity *Cianba* is a *country*. So here, point of interest is to find the noun phrases containing extracted location entity and then to identify a link between locative noun phrase and motion event trigger to identify whether this locative entity is part of the travel of *Marco Polo* or not. Example 5.8 denotes the locative noun phrase "the province of Cathay" and location entity "Ca-cian-fu" are associated with motion event trigger "*come*" and hence, they can classified as a locations visited by *Marco Polo*.

(Example 5.8) *"At the end of the four days you come to Ca-cian-fu,a large and noble city, lying to the south, in the province of Cathay."*

This relationship is examined using shallow parsing technique. In addition, it may not be a possibility to link a complex locative noun phrase directly with the identified motion verb. Many times a noun phrase is embedded in a prepositional phrase, such as **"in the city of Changu"** where *"the city of Changu"* is the noun phrase that is linked to the preposition *"in"* and this prepositional phrase **"in the city of Changu"** is linked to a motion event trigger "come". Algorithm 2 shows the method for identifying relationship between motion event trigger and directly or indirectly linked locative noun phrase:

Algorithm 2: Linking motion verbs and location entities

Input: an input sentence(S), A list of identified motion verbs(V), A list of NER identified locations(L)

Output: A list of locative noun phrases associated with motion verbs

```
/* A function returns the phrases of a phraseType for a given
   sentence                                                    */
parseTreesentence,phraseType phraseList ← {}
tree ← parse(sentence)
/* If subtree label matches the given phraseType,convert subtree to
   text                                                        */
for subtree ∈ tree do
    if subtree.label ∈ phraseType then
        phrase ← joinLeaves(subtree)
        phraseList ← phraseList ⋃ phrase
    end
end
return phraseList
/* Finding all verb phrases of a sentence                     */
verbPhrases ← parseTree(sentence, VP)
for MotionVerb ∈ MotionVerbList do
    /* Finding verb phrases only associated with motion verb  */
    for phrase ∈ verbPhrases do
        if phrase.startsWith(MotionVerb) then
            verbSpecificPhrases ← verbSpecificPhrases ⋃ phrase
        end
    end
    /* Removing overlapping verb phrases and keeping the longest
       phrase                                                  */
    longestVerbPhrase ← MAXLEN(verbSpecificPhrases)
        /* Finding noun phrases and removing overlapping noun phrases
        */
    nounPhrases ← parseTree(longestVerbPhrase, NP)
     longestNounPhrases ← maxlen(nounPhrases)
    for NP ∈ longestNounPhrases do
        nounPhraseTokens ← FindTokens(longestNounPhrase)
            /* If the noun phrase contains the NER identified location,
            mark the noun phrase as travelled noun phrase       */
        for location ∈ NERLocations do
            if location ∈ nounTokens then
                travelledNounPhrases ←
                 travelledNounPhrases ⋃ NP
            end
        end
    end
end
```

Algorithm 2 parses the sentence to identify all possible verb phrases and keeps the verb phrases that are linked to motion event triggers. Then it checks the structure of a verb phrase. A verb phrase structure is as below:

VP-> VBD NP* PP* OR VP-> VBD PP* NP*

Here, VP represents verb phrase, NP represents noun phrase and PP represents prepositional phrase. And * sign denotes any number of phrases. Now there could be various path features in identified VP as shown below Table 5.5.

Table 5.5: FrameNet Frames for example verb

Different structures of VP
VB VP PP
VB VP PP NP
VB VP NP
VB VP NP PP
VB VP ADVP NP PP

Now the algorithm identifies the structure of the verb phrase. If the prepositional phrase is associated with motion event trigger, then it is selected and this prepositional phrase is again broke down into a tree structure until all possible noun phrases associated with motion event trigger are found. If any of these noun phrases contain the extracted locative entity, then the locative entity is marked as a location visited by Marco Polo.

So as a summary, location entity was identified from the narrative of *Marco Polo* using various Named Entity Recognition techniques. In the next step, verbs were considered as a motion event triggers and various lexical resources were used to classify verbs as motion verbs and non motion verbs. After identifying all possible motion verbs, their context was checked to determine whether they should be marked as a motion event triggers or not. In the next step, direct or indirect relationship between motion event triggers and locative noun phrase containing location entities are examined and the locations are marked as a location visited of *Marco Polo* if locative entity and motion event trigger are linked. As discussed in Chapter 1 of Introduction, the sequence of travel path within any part of the narrative is in the order of actual travel path of *Marco Polo* and order of Part 1 and Part 2 is interchanged after mutual examination of the travel path. So all the locations that were marked as a location visited by *Marco Polo*, are chronologically connected in order to find a travel path.

Chapter 6

Results and Discussion

6.1 Evaluating Entity Extraction

The performance of NER systems is generally evaluated in terms of precision, recall and F1 score. Precision for NER systems can be defined as a percentage of correctly identified instances (of a particular entity type) with respect to all the instances (for a particular entity type). Whereas recall for Named Entity Recognition systems is, the percentage of correctly identified instances (of a particular entity type) with respect to a total absolute true number of instances (of a particular entity type). F1 score is a combination of precision and recall and does the equal balancing between them by computing the harmonic mean of precision and recall.

Now precision and recall are typically computed by measuring the entities against each other. NER systems can identify entities either fully or partially as shown in the below Example 6.1 with respect to the ground truth as shown in Example 6.2. In both the examples, entities are shown in [ ] brackets. The ground truth has three entities and NER tool used here is Stanford NER tool which identifies two entities out of three entities shown in ground truth.

(Example 6.1) *"[Marco Polo] noted that, [Armenia the Greater] is a large country and at the entrance of it is a city called [Arzinga]. "*

(Example 6.2) *"[Marco Polo] noted that, [Armenia] the Greater is a large country and at the entrance of it is a city called Arzinga."*

Here, an exact match for "Marco Polo" was found, "Arzinga" was missed and a only sub string match for "Armenia the Greater" was found. Now if the exact match is chosen as a the matching criteria then it affects both the false positive (Armenia being identified as false positive) and false negative

(Armenia the Greater being identified as false negative) as shown in Table 6.1.

Table 6.1: NER Example Results

Actual Entity	Predicted Entity	True Positive	False Positive	False Negative	True Negative
Marco Polo	Marco Polo	1	0	0	0
Armenia the Greater	Armenia	0	1	1	0
Arzinga	-	0	0	1	0

If the overlapping is allowed for the matching criteria then false positive and false negative will not be affected but "Armenia" will be counted as a true positive. But this evaluation approach can lead to entity ambiguation and it often requires human intervention to decide which overlap should be marked correctly and which overlap should be not. So to avoid that issue, exact match for entities is considered for evaluation.

Also, the mentioned metrics are computed based on the individual entity references in the text. So if a certain location is mentioned several times in a text, all the instances should be identified by NER systems as an individual reference. That means, all the different instances of "India" in the text will be taken into consideration in the calculation of precision, recall and F1.

As discussed in gazetteer preparation in Section 4.2, three different versions of gazetteers are created in this thesis but as expected, combination of Murray and Yule indexes covers more entities then each of them individually and hence, gazetteer created from combined indexes outperforms the other two as shown in results in Figure 6.1, Figure 6.2 and Figure 6.3 for Part 1, Part 2 and Part 3 respectively. So to compare with other pre-trained NER tools, the gazetteer generated from combined indexes is used and now onwards, it will be referred as a *gazetteer* in all future references.

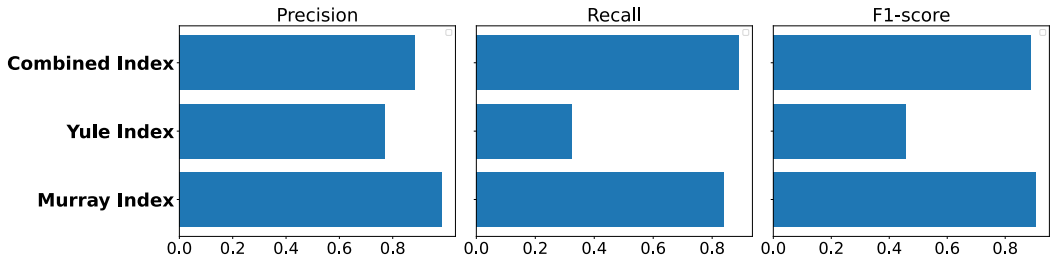

Figure 6.1: Different gazetteers comparison for PART 1

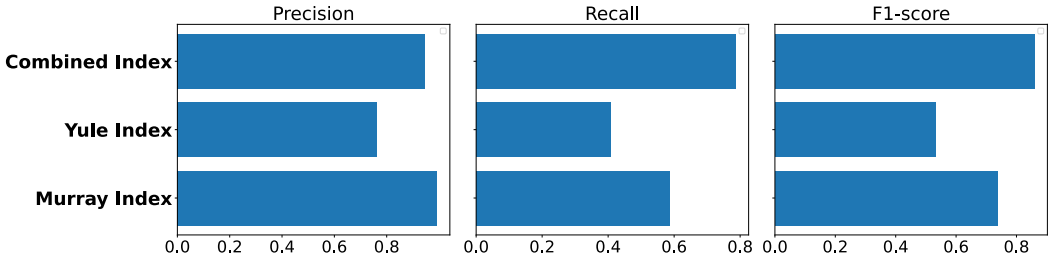

Figure 6.2: Different gazetteers comparison for PART 2

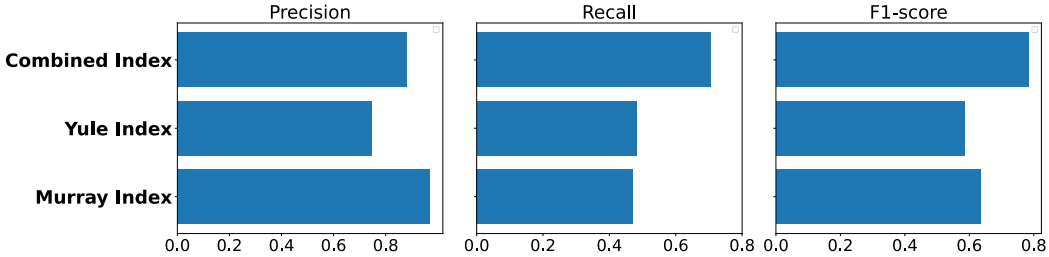

Figure 6.3: Different gazetteers comparison for PART 3

Figure 6.4, Figure 6.5 and Figure 6.6 shows the performance of different pre-trained Named Entity Recognition systems described in Section 5.1 on Part 1, Part 2 and Part 3 respectively. Figure 6.7 shows the comparison of NER tools for the entire travelogue. Gazetteer outperforms all the other pre-trained NER tools for all the all three parts of the travelogue in terms of F1-score. AllenNLP NER has almost identical precision as gazetteer but recall is significantly lower than gazetteer and hence, F1-score performance becomes slightly lower than Gazetteer but it still outperforms other pre-trained Named Entity Recognition taggers like NLTK, spaCy and Stanford. spaCy NER performs worst with F1-score of 0.47 whereas Stanford NER and NLTK NER have slightly better score than spaCy NER.

It is observed that these pre-trained NER models do not perform well on entities with different naming conventions and many of the location entities with ' - ' are missed. Another problem observed that affects the performance of pre-trained NER systems is incorrectly identified entity type. So for example, when an LOC (location) entity is identified as a PERSON, it results into a false negative for LOC and a false positive for PERSON. This problem is one of the contributing factor that affects the performance of the NER tools used in this thesis. The historical travelogue used in this thesis contains many entities that are unknown to NER tools and hence, results into false identification.

As gazetteer performs really well for the identification of location entities and achieves F1-score of 0.85. So for the subsequent tasks, location entities extracted by gazetteer is used.

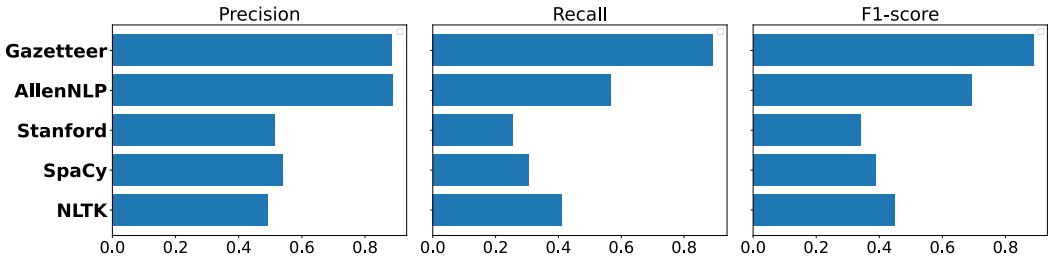

Figure 6.4: Location entity identification results for PART 1

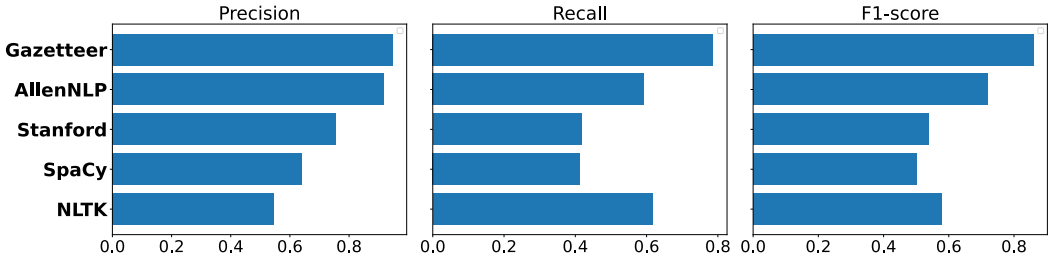

Figure 6.5: Location entity identification results for PART 2

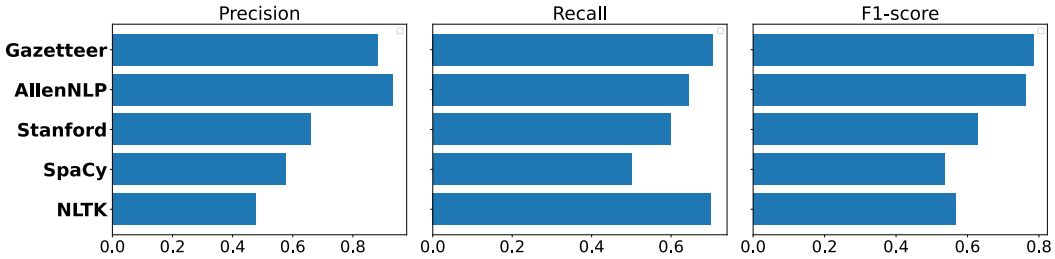

Figure 6.6: Location entity identification results for PART 3

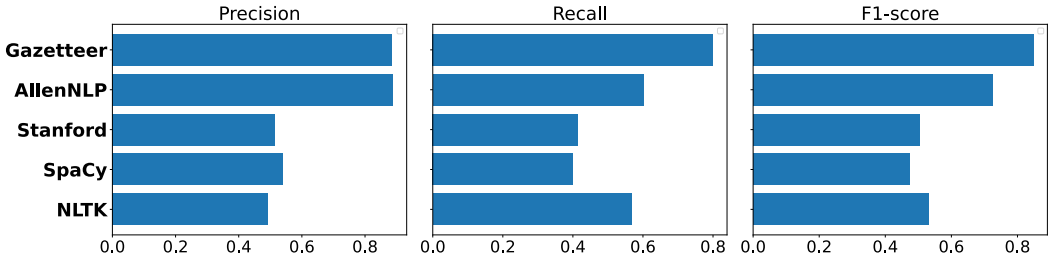

Figure 6.7: Location entity identification results for entire travelogue of Marco Polo

6.2 Evaluating Motion Events Detection

To identify the locations visited by *Marco Polo*, it is important to identify events related to motion. The motion event is detected from the narrative as described in Section 5.2 using three different lexical resources, namely, WordNet, VerbNet and FrameNet.

The comparison of these three linguistic resources is shown as below in Figure 6.8, Figure 6.9 and Figure 6.10 for Part 1, Part 2 and Part 3 respectively. Figure 6.11 shows the results for the entire travelogue of *Marco Polo*.

From the results, it can be clearly seen that all three linguistics resources follows the same pattern of having a poor precision but high recall. Lower precision reflects that the algorithm is returning more irrelevant results compare to the relevant ones. Whereas high recall represent that an algorithm is returning most of the relevant results. F1 score is a function of a precision and recall and hence, it is getting affected by a low precision. Now there is a noticeable difference between F1 score of all three resources with VerbNet having a highest and FrameNet being the lowest. This is because they have a significant difference in precision but recall has only marginal difference.

So it can be concluded that all these lexical resources are able to find the motion events triggers but they also falsely identify other events as a motion event trigger. The possible reason for low precision, that is, identification of irrelevant motion triggers can be contributed to the complexity of the natural language. Here, it depends not only the choice of the seed for all lexical resources but also on the context as well. Here, when context of motion is added as described in Section 5.2, it also adds the various other context that can be relevant. Still, the choice of seed is the most influencing factor that affects the performance. Seeds for all three lexical resources represents very broad range and it can be narrowed down to improve precision. But narrowing down the choices for seeds require in-depth background knowledge and as the goal of thesis is to automate path retracing as much as possible, choice for seeds is not narrowed down intentionally.

So here, recall seems to be a better evaluation metric and high recall is desirable because it shows less number of false negatives or simply it means that, any possible motion event trigger is not being missed. From the charts, it can be seen that VerbNet achieves high precision compare to other two but it has a lowest recall compare to other lexical resources. Where recall is high for all three lexical resources and for each part of the narrative but WordNet has a highest recall compare to VerbNet and FrameNet for each part.

So as WordNet has a higher recall of 0.80, it shows that it successfully identifies most of the motion events and hence, motion event triggers identified by WordNet will be used for subsequent tasks.

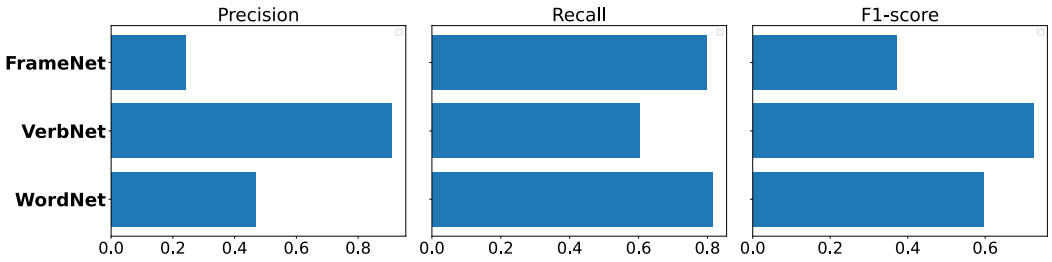

Figure 6.8: Motion event detection for PART 1

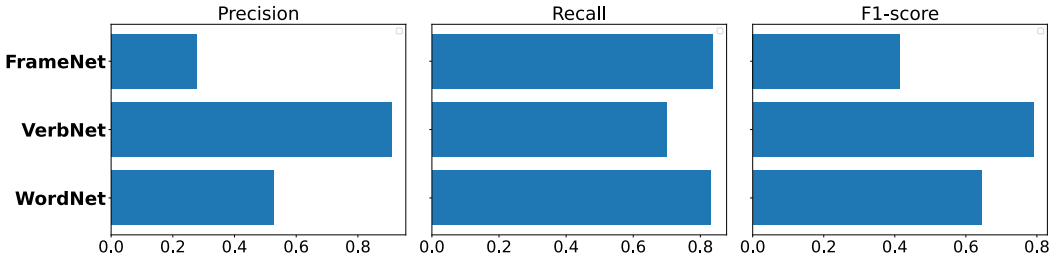

Figure 6.9: Motion event detection for PART 2

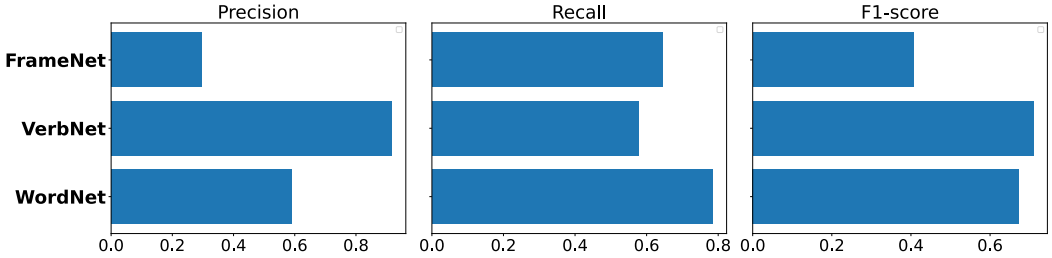

Figure 6.10: Motion event detection for PART 3

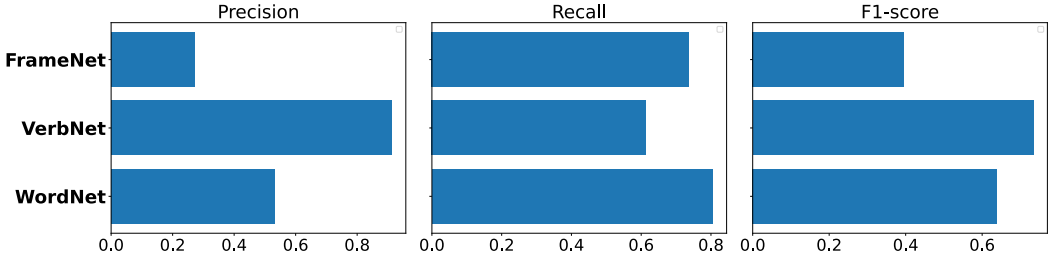

Figure 6.11: Motion event detection for entire travelogue of Marco Polo

6.3 Evaluating motion Events and Location Entities Linking

To interpret the meaning of a sentence and to identify whether it describes any information about the journey of Marco Polo or the places he visited, various different important information from the sentences have been extracted independently of each other for the entire travelogue of *Marco Polo*. At this stage, location entities and motion event triggers are identified and now it is important to examine the link between them. A relationship between verbs and entities can be understood by extracting noun phrases that are linked to motion verbs. Sometimes noun phrases are directly linked with a motion verb and sometimes they are embedded in prepositional phrases and those prepositional phrases are linked to motion verbs. The goal here is to identify verb phrase related to motion event triggers and then to extract all possible noun phrase linked with it directly or indirectly using the algorithm described in Section 5.3. Out of these all noun phrases, locative expressions are identified using the extracted location entity. If any of the associated noun phrase exist with extracted location entity then location entity and motion event trigger are linked and location entity is marked as a location visited by *Marco Polo*. This can be demonstrated with an example sentence given in Example 6.3. This sentence contains three possible motion triggers, namely *leaving, proceeding, reach.*.

(**Example 6.3**) *"Leaving the city of Sa-yan-fu, and proceeding fifteen days journey towards the south-east, you reach the city of Sin-gui, which, although not large, is a place of great commerce."*

The relationship between motion event trigger and its associated locomotive expression (in the form of noun phrase) for each motion verb can be shown as below:

For verb **leaving**,
"Leaving the city of Sa-yan-fu, and proceeding fifteen days journey towards the south-east, you reach the city of Sin-gui, which, although not large, is a place of great commerce."

For verb **proceeding,**
"Leaving the city of Sa-yan-fu, and proceeding fifteen days journey towards the south-east, you reach the city of Sin-gui, which, although not large, is a place of great commerce."

For verb **reach**,
"Leaving the city of Sa-yan-fu, and proceeding fifteen days journey towards the south-east, you reach the city of Sin-gui, which, although not large, is a place of great commerce."

WordNet has a highest F1-score than other two for the entire travelogue of Marco Polo. So for retracing the travel path of *Marco Polo*, results of gazetteer for location identification and results of WordNet for motion event triggers are considered. These two information is then linked as discussed above to identify the locations that are actually visited by Marco Polo.

Here, identification of all possible traveled location entities along with correct identification of traveled locations is important and hence, precision and recall both are important evaluation metrics so we can judge the performance of an algorithm by precision, recall and F1 score. Figure 6.12, 6.13 and 6.14 shows the results of traveled location identification for Part 1, Part 2 and Part 3 respectively and Figure 6.15 shows the results for entire travelogue. It can be concluded that recall is very low for all three parts of the book. It means that the method presented here identifies less number of traveled locations. But the precision for all three parts of the book is very high and it denotes the accuracy of an approach.

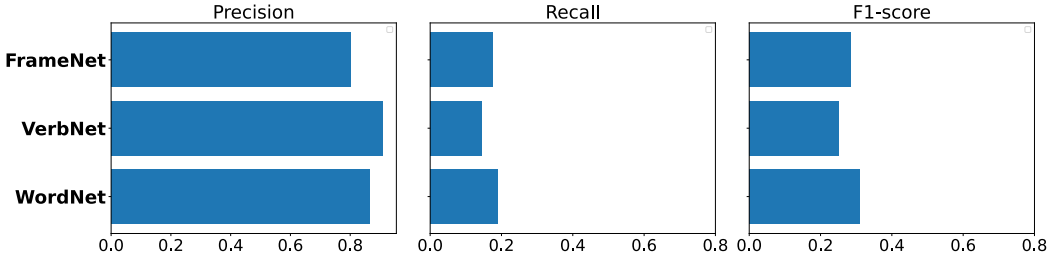

Figure 6.12: Travelled location identification results for PART 1

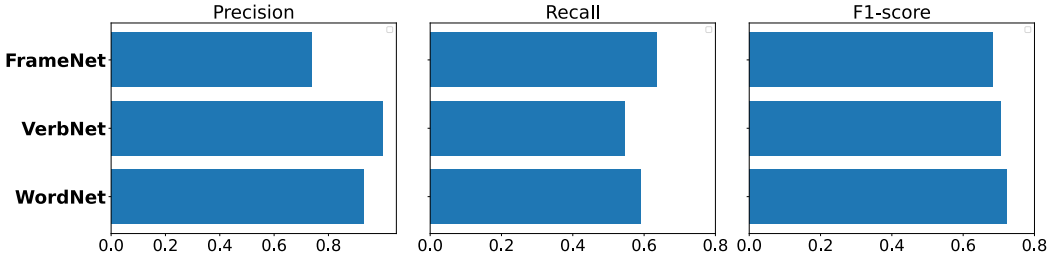

Figure 6.13: Travelled location identification results for PART 2

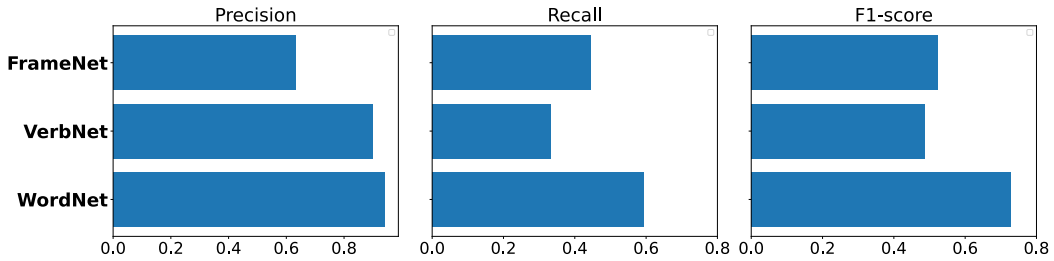

Figure 6.14: Travelled location identification results for PART 3

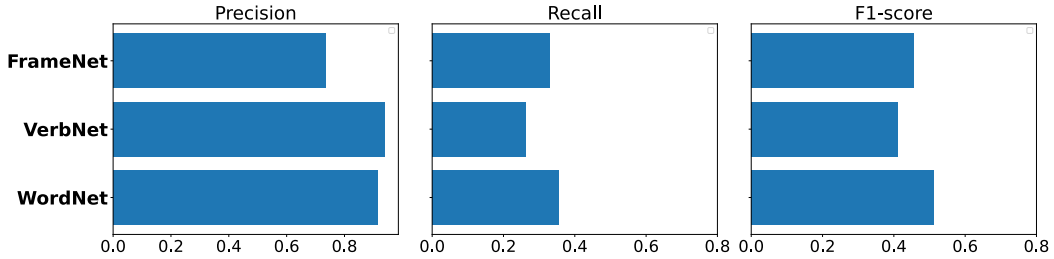

Figure 6.15: Travelled location identification results for entire travelogue

Here, from the results, it can be seen that recall and hence, F1-score for Part 1 is remarkably low. For Part 2 and Part 3, the results are better than Part 1 but still the results are poor. This limitation of an approach is that it cannot do multiple relation extraction and verbs linking from single independent sentences and it can be shown via below Example 6.4.

(Example 6.4) "***Leaving*** *Ta-in-fu, and* ***riding*** *westward full seven days through very fine districts, amid numerous merchants, you* ***find*** *a large town,* ***named*** *Pi-an-fu,supported by commerce and the silk manufacture."*

So here, the approach described above, marks the location *Ta-in-fu* as a visited location but it fails to identify the place *Pi-an-fu* as a location visited because *Pi-an-fu* is linked to verbs 'named' and 'find' and these two are not motion event triggers. Here, verbs 'find' and 'named' should get linked to a verb 'riding'. 'Riding' is a verb that denotes the event of motion, going to a city named 'Pi-an-fu' and a link between these three verbs can be easily understood by reading the sentence but the approach mentioned here fails to identify that.

As discussed in Chapter 1, these identified traveled locations are then mapped to their contemporary equivalents manually by the author of this thesis and connected in a chronological order as discussed in Chapter 1 to generate a travel path. This generated travel path is as shown in Figure 6.16.

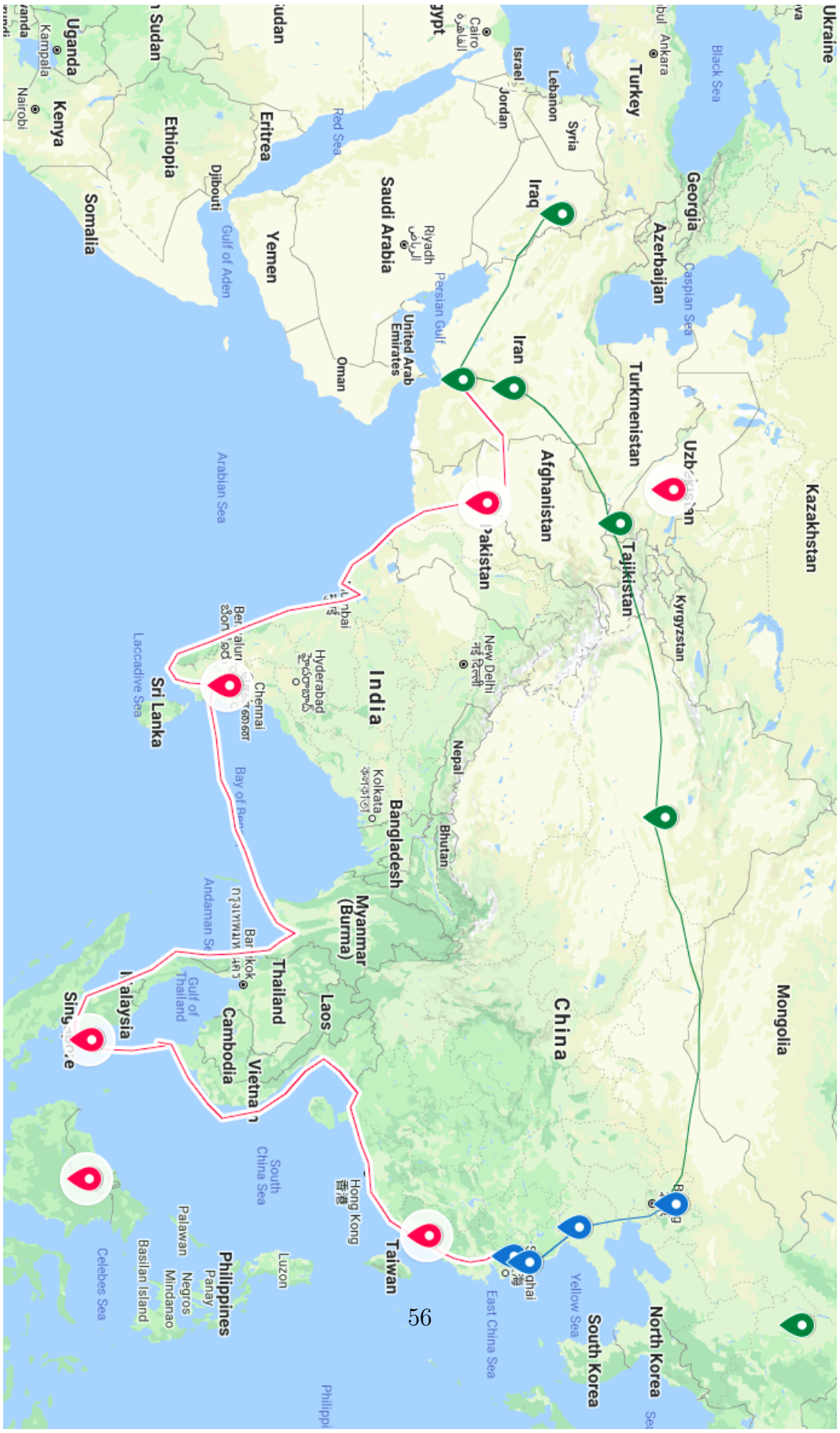

Figure 6.16: Identifed travel path of Marco Polo

Now from the map, it can be seen that, the path of return journey of *Marco Polo* from China to Venice is incomplete. The reason is, many times in the entire travelogue, specially, on his return journey, *Marco Polo* talks about traveled places using words that do not denote an event of motion like Example 6.5 and Example 6.6. The narrative never explicitly states that *Marco Polo* returned to Great Turkey or Venice but it can be inferred by reading Example 5.5 and Example 6.6, and from the domain knowledge.

(Example 6.5) *"Having told you all about these Tartars of the East, I might go on to* ***treat*** *of* ***Great Turkey;"***

(Example 6.6) *"But I believe it was the pleasure of God,.....,there never was a man,.....,who explored so much of the world as did Marco, the son of Nicolo Polo, that noble and great citizen of* ***Venice****."*

Another observation is that, that narrative itself is very ambiguous. From the narrative itself, it is a very difficult to separate the locations *Marco Polo* actually visited from the locations he is just talking about. Consider the sentences given in Example 6.7 and Example 6.8. These both sentences have the same sentence structure but in Example 6.7, the place reference *Zanghibar* is a location in *East Africa* and *Marco Polo* never went there. He is describing it from the stories he heard about from other people. Whereas in Example 6.8, the place reference *Balk* is an important location on *Marco Polo* route and it is the location that he actually visited.

(Example 6.7) *"Having nothing more to tell of this island, I will* ***go on*** *to that of Zanghibar."*

(Example 6.8) *"Now I will* ***go on*** *to another city named Balk."*

Because of above reasons, there is a difference of opinion between researchers about the exact travel path of *Marco Polo* and there has not been single agreed and verified path.

Chapter 7

Conclusion

7.1 Conclusion

This thesis presents Natural Language Processing based approach to retrace the travel path taken by Marco Polo from his 12th century old narrative. In addition, it presents various gold standard annotations set ups that are created manually and that can be used to evaluate the performance of NLP techniques at the various stages using relevant evaluation metrics. It also describes the challenges involved with historical datasets as well as challenges associated with domain-specific travelogue of Marco Polo.

To separate the stories related to traveling of Marco Polo from the rest of the stories, an approach is presented to identify events that describe motion using various lexical resources such as WordNet, VerbNet and FrameNet. These all lexical resources follow the pattern of high recall but low precision. Simply, these lexical resources identify instances of motion events correctly but they also identify other events as motion events.

In addition, location entities are detected using various state-of-the-are Named Entity Recognition techniques such as NLTK NER, spaCy NER, Stanford NER and AllenNLP NER, and external knowledge from the different gazetteers. It can be concluded that gazetteer performs well compare to pre-trained state-of-the-art NER techniques except AllenNLP NER which has slightly low performance compare to gazetteer. State-of-the-art techniques do not work well with historical texts which has a different naming conventions. They fail to identify entities with special characters or misclassify their entity type.

After that, link between motion event triggers to locative expressions is examined to identify which locative expressions are part of the journey and which locative expressions are part of the other stories in the narrative. Locations that are explicitly mentioned as visited are identified but those locations

that are not explicitly described as visited were vague and explained using narrative event and not motion event. At the end, the identified locations that are visited by Marco Polo, are mapped to current day locations and connected chronologically to identify the travel path and this path is visualized on a map.

Out of several, one of the major problem encountered in this thesis was, not a clear distinction between narrative of the places Marco Polo visited in reality vs the narrative of the places that Marco Polo only describes. And hence, there is a difference of an opinion among researches about the exact path taken by Marco Polo and hence, the final evaluation depends on the manual gold standard setup that was done manually as per the understanding of the author of the thesis.

7.2 Future Work

This thesis presented a detailed look for the task of extracting travel path from the historical narrative of *Marco Polo*. It then also presented the challenges with historical dataset and in depth analysis quantitative and qualitative analysis of the narrative. However, certain aspects and improvements in the approach are yet to be explored for improving the accuracy of an extracted travel path. This final chapter introduces the aspects that can be further explored and generalize the process of separating various themes or stories from the historical texts.

One of the most important task is to extract the location entities. Now from the presented results, it can be seen that the various state-of-the-art tools like spaCy NER and Stanford NER have poor performance for the task of location entity extraction on a historical narrative. The main reasons are different naming conventions of historical corpora and misclassification of entities. Hence, further research is needed to explore the underlying reasons behind the poor performance and how it can be improved so the need for using domain-specific gazetteer can be eliminated.

In this thesis, motion event triggers are identified using strictly motion related seeds. However, throughout the entire narrative, an event of motion is described indirectly many times using other events so further research is needed to identify instances of motion events that are not explicitly expressed. Some recent research that has been carried out already tries to reconstruct meanings [56] from the text. However, a further exploration is required to understand the underlying meaning of sentences from the historical narratives.

Another problem encountered in thesis is linking location entities and motion event triggers for long sentences because of the complexity of a natural language. A further research is helpful for multiple relation extraction and

verbs linking from single independent sentences.

In addition, a further research is needed to determine the order of the travel. In the thesis, extracted locations are connected in a chronological order to reconstruct the travel path but it requires in-depth domain knowledge to understand that narrative itself is written in a chronological order and hence, extracted locations can be connected in a chronological order. But a research is needed that can find the order of extracted locations.

An interesting experiment that can be performed on this narrative is to develop an automatic generalized approach for path extraction that can also be applied to other characters in the narrative to draw their trajectories. The reason behind that is, this narrative contains many important information about the historical times and other famous characters and it can extracted using NLP as well.

Appendix A

Part-of-Speech Tags

Table A.1: Part-of-Speech Tags

Type	Description
CC	Conjunction, coordinating
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Conjunction, subordinating or preposition
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Verb, modal auxillary
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Noun, proper singular
NNPS	Noun, proper plural
PDT	Predeterminer
PRP	Pronoun, personal
PRP$	Pronoun, possessive
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Adverb, particle
SYM	Symbol
TO	Infinitival to
UH	Interjection
VB	Verb, base form
VBZ	Verb, 3rd person singular present
VBP	Verb, non-3rd person singular present
VBD	Verb, past tense
VBN	Verb, past participle
VBG	Verb, gerund or present participle
WDT	wh-determiner
WP	wh-pronoun, personal
WP$	wh-pronoun, possessive
WRB	wh-adverb
#	Pound sign
$	Dollar sign
.	Sentence final punctuation
,	Punctuation mark, comma
:	Punctuation mark, colon
(	Left bracket character
)	Right bracket character
"	Straight double quote

Appendix B

Chunk Tags

Table B.1: Types of Chunk Tags

Type	**Description**
NP	Noun phrase
VP	Verb phrase
PP	Prepositional phrase
ADVP	Adverb phrase
ADJP	Adjective phrase
SBAR	Subordinating conjunction
PRT	Particle
INTJ	Interjection

Bibliography

[1] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, 1998.

[2] Adrien Barbaresi. A constellation and a rhizome: two studies on toponyms in literary texts. *VISUALISIERUNG*, page 167, 2018.

[3] Adrien Barbaresi. Placenames analysis in historical texts: tools, risks and side effects. In *Corpus-based Research in the Humanities*, 2018.

[4] Fabian Barteld. Detecting spelling variants in non-standard texts. In *Proceedings of the student research workshop at the 15th conference of the European chapter of the association for computational linguistics*, pages 11–22, 2017.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[6] Marcel Bollmann. A large-scale comparison of historical text normalization systems. *arXiv preprint arXiv:1904.02036*, 2019.

[7] Lars Borin and Markus Forsberg. Something old, something new: A computational morphological description of old swedish. In *LREC 2008 workshop on language technology for cultural heritage data (LaTeCH 2008)*, pages 9–16, 2008.

[8] Lars Borin, Dimitrios Kokkinakis, and Leif-Jöran Olsson. Naming the past: Named entity and animacy recognition in 19th century swedish literature. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007).*, pages 1–8, 2007.

[9] Loes Braun, Floris Wiesman, and Ida Sprinkhuizen-Kuyper. Information retrieval from historical corpora. In *Proceedings of the 3rd Dutch Information Retrieval Workshop (DIR2002)*, pages 106–112, 2002.

[10] Marco Dinarelli and Sophie Rosset. Tree-structured named entity recognition on ocr data: Analysis, processing and results. 2012.

[11] Maud Ehrmann, Giovanni Colavizza, Yannick Rochat, and Frédéric Kaplan. Diachronic evaluation of ner systems on old newspapers. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, number CONF, pages 97–107. Bochumer Linguistische Arbeitsberichte, 2016.

[12] Eugenia Eumeridou, Blaise Nkwenti-Azeh, and John McNaught. An analysis of verb subcategorization frames in three special language corpora with a view towards automatic term recognition. *Computers and the Humanities*, 38(1):37–60, 2004.

[13] Peter Exner and Pierre Nugues. Using semantic role labeling to extract events from wikipedia. In *DeRiVE@ ISWC*, pages 38–47, 2011.

[14] Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACLâ05)*, pages 363–370, 2005.

[15] Antske Fokkens, Serge Ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, et al. Biographynet: Methodological issues when nlp supports historical research. In *LREC*, pages 3728–3735, 2014.

[16] Roger Garside. A hybrid grammatical tagger: Claws 4. *Corpus annotation: Linguistic information from computer text corpora*, 1997.

[17] J Gerhard and W van den Heuvel. Survey report on digitisation in european cultural heritage institutions 2015. Technical report, Technical report, Europeana/ENUMERATE, June, 2015.

[18] Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288, 2002.

[19] Rafael Giusti, Arnaldo Candido Jr, Marcelo Muniz, Lívia Cucatto, and SM Aluísio. Automatic detection of spelling variation in historical corpus: An application to build a brazilian portuguese spelling variants dictionary. In *Corpus Linguistics*, 2007.

[20] Michael F Goodchild and Linda L Hill. Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039–1044, 2008.

[21] Claire Grover, Sharon Givon, Richard Tobin, and Julian Ball. Named entity recognition for digitised historical texts. In *LREC*, 2008.

[22] Christian Hardmeier. A neural model for part-of-speech tagging in historical texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 922–931, 2016.

[23] Iris Hendrickx, Michel Généreux, and Rita Marquilhas. Automatic pragmatic text segmentation of historical letters. In *Language Technology for Cultural Heritage*, pages 135–153. Springer, 2011.

[24] Annette Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9(3):341–378, 1985.

[25] M Honnibal and I SpaCy Montani. 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. 2017.

[26] Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare Voss, Jiawei Han, and Avirup Sil. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–268, 2016.

[27] Paul Kingsbury, Martha Palmer, and Mitch Marcus. Adding semantic annotation to the penn treebank. In *Proceedings of the human language technology conference*, pages 252–256. San Diego, California, 2002.

[28] Paul R Kingsbury and Martha Palmer. From treebank to propbank. In *LREC*, pages 1989–1993. Citeseer, 2002.

[29] Karin Kipper, Hoa Trang Dang, Martha Palmer, et al. Class-based construction of a verb lexicon. *AAAI/IAAI*, 691:696, 2000.

[30] Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. A hybrid approach for event extraction and event actor identification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 592–597, 2011.

[31] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[32] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*, 2016.

[33] Beth Levin. *English verb classes and alternations: A preliminary investigation*. University of Chicago press, 1993.

[34] Johann-Mattis List, Simon J Greenhill, and Russell D Gray. The potential of automatic word comparison for historical linguistics. *PloS one*, 12(1):e0170046, 2017.

[35] Fei Liu, Maria Vasardani, and Timothy Baldwin. Automatic identification of locative expressions from social media text: A comparative analysis. In *Proceedings of the 4th International Workshop on Location and the Web*, pages 9–16, 2014.

[36] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[37] Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. 1993.

[38] Katherine McDonough, Ludovic Moncla, and Matje van de Camp. Named entity recognition goes to old regime france: geographic text analysis for early modern french corpora. *International Journal of Geographical Information Science*, 33(12):2498–2522, 2019.

[39] Marcia Munoz, Vasin Punyakanok, Dan Roth, and Dav Zimak. A learning approach to shallow parsing. *arXiv preprint cs/0008022*, 2000.

[40] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[41] Patrick Olivier and Junâichi Tsujii. A computational view of the cognitive semantics of spatial prepositions. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 303–309, 1994.

[42] Sharon Myrtle Paradesi. Geotagging tweets using their content. In *Twenty-Fourth International FLAIRS Conference*, 2011.

[43] Roger L Payne. Development and implementation of the national geographic names database. *Names*, 43(4):307–314, 1995.

[44] Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*, 2017.

[45] Michael Piotrowski. Natural language processing for historical texts. *Synthesis lectures on human language technologies*, 5(2):1–157, 2012.

[46] Paolo Plini, Sabina Di Franco, and Rosamaria Salvatori. One name one place? dealing with toponyms in wwi. *GeoJournal*, 83(1):87–99, 2018.

[47] Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H Martin, and Dan Jurafsky. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 233–240, 2004.

[48] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155, 2009.

[49] Paul Rayson, Dawn E Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. Tagging the bard: Evaluating the accuracy of a modern pos tagger on early modern english corpora. In *Proceedings of the Corpus Linguistics conference: CL2007*, 2007.

[50] Ludwig Richter, Johanna Geiß, Andreas Spitz, and Michael Gertz. Heidelplace: An extensible framework for geoparsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 85–90, 2017.

[51] Kepa Joseba Rodriquez, Mike Bryant, Tobias Blanke, and Magdalena Luszczynska. Comparison of named entity recognition tools for raw ocr text. In *Konvens*, pages 410–414, 2012.

[52] Erik Tjong Kim Sang. Improving part-of-speech tagging of historical text by first translating to modern text. In *International Workshop on Computational History and Data-Driven Humanities*, pages 54–64. Springer, 2016.

[53] Humphrey Southall, Ruth Mostern, and Merrick Lex Berman. On historical gazetteers. *International Journal of Humanities and Arts Computing*, 5(2):127–145, 2011.

[54] Rachele Sprugnoli and Sara Tonelli. Novel event detection and classification for historical texts. *Computational Linguistics*, 45(2):229–265, 2019.

[55] Madhuri A Tayal, Mukesh M Raghuwanshi, and Latesh Malik. Word net based method for determining semantic sentence similarity through various word senses. In *Proceedings of the 11th international conference on natural language processing*, pages 139–145, 2014.

[56] Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. (re) construing meaning in nlp. *arXiv preprint arXiv:2005.09099*, 2020.

[57] Seth Van Hooland, Max De Wilde, Ruben Verborgh, Thomas Steiner, and Rik Van de Walle. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities*, 30(2):262–279, 2015.

[58] Atro Voutilainen. Part-of-speech tagging. *The Oxford handbook of computational linguistics*, pages 219–232, 2003.

[59] Miguel Won, Patricia Murrieta-Flores, and Bruno Martins. ensemble named entity recognition (ner): evaluating ner tools in the identification of place names in historical corpora. *Frontiers in Digital Humanities*, 5:2, 2018.

[60] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*, 1994.

[61] Yi Yang and Jacob Eisenstein. Part-of-speech tagging for historical english. *arXiv preprint arXiv:1603.03144*, 2016.

[62] Jordan Zlatev. Spatial semantics. In *The Oxford handbook of cognitive linguistics*. 2007.