Leipzig University
Institute of Computer Science
Degree Programme Computer Science, B.Sc.

Mapping the Travel Routes of Marco Polo

Bachelor's Thesis

Yiwen Cao

1. Referee: Dr. Andreas Niekler
2. Referee: Dr. Magdalena Anna Wolska

Submission date: March 28, 2022

Declaration

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

Leipzig, March 28, 2022

. . . .
Yiwen Cao

Abstract

Nowadays Natural Language Processing Models have achieved good accuracy in many tasks, like Named Entity Recognition. But the usage of Natural Language Processing techniques in historical texts is still a small domain. This thesis tries to discuss the possibility and difficulty of using Natural Language Processing techniques in historical text, by reconstructing the travel route of Marco Polo from his historical travelogue, Named Entity Recognition and thematic (movement) verb extraction are used and evaluated. Several gold standard corpus were built as well.

Contents

Chapter 1

Introduction

1.1 Motivation

Nowadays, more and more ancient literature and documents are being digitized. Despite the fact that digitization makes it more straightforward for specialists to get to sources, data analysing from the literature is still time-consuming and labor-intensive. Natural language processing has achieved relatively good accuracy rates in contemporary texts, but its use in historical texts is still in the minority. For both humanists and computer scientists, extracting and analyzing information from historical text to generate standardized, structured databases with less human work remains a challenge.

The goal of this thesis is to extract the travel route of *Marco Polo* from the famous 13th-century Work, *The Travels of Marco Polo*. In this process, some state-of-the-art tools and models, that are commonly used in Natural Language Processing, will be used. Their performances in extracting information from historical texts and classifying will be compared with human understanding, as well as some of the difficulties and conclusions reached in the process will be interpreted and analysed.

As an observation, we see that the data we used in this thesis was published in the early 20th Century, which indicates that the syntax of the text is still in quite modern English. So that some of the Nature Language Processing basic grammar analysis tools, for example Part-of-Speech Tags, are still reliable. But the translators kept the raw names of the entities in the context, there are still specific observations found in the research.

Named Entity Recognition and Ambiguities Resolving in Historical Context The Named Entities can not be recognized directly by the models. Human reorganization also needs specific knowledge. As shown in Table 1.1, the location names Marco Polo used in the book, is greatly different from the modern names, which are commonly used nowadays. Sometimes a location

Table 1.1: Examples of the Location names used by Marco Polo and the Modern Names

Polo's Name	Modern Name
Paipurth	Bayburt
Cacanfu	Hejianfu
Sinju	Xining

may be even considered as other locations, for example, *'Hormus'* is considered as a city on the Hormuz Island nowadays, but should be refers to the city on the mainland at Marco Polo's time, which now has the name *'Minab'*. The huge differences prevent the models used to identify place names from working properly. Even the data got from the models can not be used directly. Extra knowledge is needed to link the name from the book to the real place on the map.

This Situation also exists on other Named Entities. Like the title of Tartar Lord, *'Khan'*, is written as *'Kaan'* in the book, recognized as none named entity by several models.

For most of the Named Entities, when they appear in the book for the first time, they are all capitalised, like the bolded part in Example 1.1.1. And when they reappear next, the initials are still capitalized, which is helpful in identifying as named entities. However, for some Product named entities, they are always in lowercase and recognized as none named entity by several models.

Motion Events Identifying Locations in the travel route are linked with travel verbs, so finding verbs which express motion, can be a fast way to identify whether the locations are in the travel routes.

In addition to the clearly named locations, the direction between the two locations and the length of the dates travelled are also specified, as the underlined parts show in Example 1.1.1. Due to the fact that modern roads are not the same as ancient roads, and since Marco Polo passed through plains, deserts, plateaus and jungles, the pace of his advance was not the same, this can only be used as a rough guide.

Example 1.1.1 *"Travelling through a succession of towns and villages that look like one continuous city, <u>two days</u> further on to the <u>south-east</u>, you find the great and fine city of **GHIUJU** which is under Kinsay."*

Location and Motion Linking in non-chronological Context As shown in the Example 1.1.2, the book does not only contain the motion events of Marco Polo's, but also of other characters in the stories he told in the book.

For a moving verb, the subject is the person who moves, and the noun phrases containing location names is often ruled, either directly or indirectly, by the verbal relationship. By identifying the subject of the motion verb, we tried to classify whether the location name attached to the moving verb is in Marco Polo's route, like the verb phrases shown in the Example 1.1.3 and 1.1.2.

Example 1.1.2 *Motion Verbs of other Characters*
"Chinghis Kaan with all his host ***arrived*** *at a vast and beautiful plain which was called Tanduc."*

Example 1.1.3 *Motion Verbs of Marco Polo*
"When you have ***gone*** *15 miles from the city of Unken, you* ***come*** *to this noble city which is the capital of the kingdom."*

The ideal situation for extracting a motion event is to get a clear origin and destination, and it is even better if the previous destination is the next origin. In Example 1.1.4, the first motion event has the orgin *Trebizond* and the destination *Tauris*, which is also the origin of the second motion event.

Example 1.1.4 *"...that you pass in going from Trebizond to Tauris...From Tauris to Persia is a journey of twelve days."*

But there are also cases in the book where, after reaching a large region, the motion event continues to a smaller region inside it. In Example 1.1.5, the first motion event mentions Marco Polo entered the province called *Manzi*. After several chapters when he kept travelling in *Manzi*, he entered a kingdom called *Fuju* which is also a part of the province of *Manzi*.

Example 1.1.5 *"...And when you pass this river you enter the great province of MANZI...The other kingdom which we now enter, called* ***Fuju,*** *is also one of the nine great divisions of Manzi..."*

1.2 Research Objectives

In order to achieve the goal of extracting travel events from the context, the process is divided into three specific steps.

1. Identifying the Named Entities in the text, especially those representing place names.

2. Identifying the motion verbs in the text and marking the travel events from all events.

3. Connecting the route undertaken by Marco Polo and the destinations of the travel events based on the subjects and objects of the motion verbs.

1.3 Contributions

Chapter 2 gives an overview of previous work on tracing route in general and in historical works. Chapter 3 introduces the resources used in this thesis to approach the goal. In chapter 4, the specific details of the dataset used for training and evaluation are introduced. In Chapter 5 the performances of the named entity recognition software packages and the event extracting algorithms are evaluated. Additionally, in the whole process, several gold standard annotated datasets were built based on domain knowledge and can be used for further research.

Chapter 2

Background

2.1 Named Entity Recognition in Historical Work

Named Entity Recognition (NER) is a sub-task of Information Extraction, which extracts the informative item names, including person names, organization names and location names, and numbers [9]. The annotations of named entities, the relationships between them and the corresponding events has gone from light to rich and the rules have been standardised[15].There are many approaches to the extraction in the current text, state-of-the-art statistical-based models have achieved good results in the current text[24][23].

The limit of traditional Knowledge-based natural language processing systems is a domain-specific dictionary that requires a great deal of manual knowledge engineering, therefore extracting information automatically to build a dictionary is tried[12]. Descriptive representations are unstructured metadata, which cannot be used directly by machines. A research based on the descriptive fields of the Smithsonian Cooper–Hewitt National Design Museum in New York explores the value and limitations of entity extraction on it[20]. By building a NER model in German, the performance of linear-chain CRFs and BiLSTMs are compared in large-scale and small-scale data scenarios. BiLSTM outperforms CRFs on large-scale data, but the opposite is true for small datasets. However, BiLSTM can be trained on multiple corpora by transfer learning[11].

NER in historical texts, on the other hand, is still a relatively small domain, but there are still a number of attempts using various methods. Five different NER models were used on two historical corpora and later consolidate into a single result by a voting system. This voting system gives consistent results for both corpora, its precision and recall outperform the individual models[21]. Furthermore the authors argued that the statistical NER approach based on supervised learning does not depend strongly on whether the corpus is preprocessed into perfect modern English.

As a important part of NER, to link the location names, i.e toponym, with the right entity is so called toponym Disambiguation. Quantitative attempts at toponym disambiguation have been made by calculating a score for each possible toponym based on the context. Methods for calculating scores are summarised in three categories: map-based representation of place names, knowledge-based, and data-driven or supervised learning[4]. Other research uses automatic recognition of toponyms by some methodes, which are independent from the language they used, makes it possible to do Toponym Disambiguation for corpora in several languages[10]. A very important gazetteer database of this concept is GeoNames. Its quality and accuracy of the data are counted, the errors arising from inaccuracies are concluded and solutions are proposed[1]. Another attempt is to combine generalized event extraction with toponym disambiguation, using other factors in the event structure for toponym disambiguation. It is proved that through events, it is possible to turn toponym disambiguation into a probabilistic problem that can be solved by logistic regression[13].

2.2 Tracing routes

In the fields of digital humanities, fieldwork on culture route has been done online[5]. Combining with linguistic analysis, geographic information extraction and visualisation, place names will be extracted from historical texts to customisable maps[3]. The geographic information system (GIS) plays a very large role in digital humanities. Advanced spatial analyses in GIS, such as Cost Surface Analysis (CSA) and Least Cost Path Analysis (LCP), are used to provide a more nuanced interpretation of more literary works, such as historical travel works and topographical documents[8].A practical example is the visualisation of two literary works[2].

In the specific area of reconstruction of Marco Polo's travel route, a detailed example of manual work by using Google Earth and Wikimapia, was made [17][16][18][19].Based on human understanding and analyzing the text, it shows how to use the new techniques to do the traditional job. Another example is the experience of using Google Earth to reconstruct the Silk Road[7].

Chapter 3

Resources

In this chapter, the resources used in this thesis are introduced. In Section 3.1, the Natural Language Processing Tools used for basic text processing and named entity recognition are introduced. Section 3.2 describes the lexical resources used to detect motion verbs. Section 3.3 introduces geographic information system systems used to identify the locations in the thesis.

3.1 Natural Language Processing Tools

Flair Flair[1] is a state-of-the-art Natural Language Processing framework based on Pytorch and built on Python. It has a simple interface to use and combines different embeddings. In this thesis, Flair is used for the named entity recognition task. The Flair NER model is based on document-level XLM-R embeddings and FLERT[14]. The model reaches a F1-score of 90.93. Fine-tuning of the model needs to make the tag dictionary from the corpus with the combination of Inside–outside–beginning tags and named entity recognition tags, then initialize fine-tuneable embeddings and sequence tagger.

NLTK NLTK[2] stands for Natural Language Toolkit. It is a powerful platform to do Natural Language Processing jobs in Python. It has a interface to more than 50 corpora and lexical resources. And it has suite of libraries for all kinds of text processing tasks. In this thesis, its interface to lexical resources and basic functions like tokenization, tagging and parsing are used.

[1] https://github.com/flairNLP/flair

[2] https://www.nltk.org

Stanford Parser Stanford Parser[3] for constituency parsing used shift-and-reduce operations to process input sentences according to the productions of grammar to a constituency parse tree. It is a bottom-down parsing experience. Pushing the next word in the input sentence on stack (shift), and if the top n items on stack match, popping them off the stack and pushing partial parse tree on stack (reduce), this process repeated until reducing to a parse tree with root node S.

3.2 Lexical Resources

VerbNet VerbNet[4] is the largest online English verb lexicon, with approximately 5800 English verbs. The verbs are grouped by similar semantics and syntax into classes and saved in a tree-structured hierarchy. In the hierarchy, each verb class under the top-level number (9-109) has a shared semantic relationship. Classes number 1-57 are developed directly from Levin's verb classes, others are developed later [6]. The classes can be specified into subclasses by manners and instruments of verbs. Each verb class is described by semantic roles, which describe the conceptual relations participants in a sentence consistently despite the change of the syntax, selectional restrictions, which further restrict the semantic roles, syntactic frames and semantic predicates. Currently 35 semantic roles are used in VerbNet. In this thesis, WordNet interface from NLTK[5] is used.

FrameNet FrameNet[6] is a English lexicon based on annotated examples of words' meaning and usage in real texts. There are more than 1,200 semantic frames with more than 13,000 word senses, linked to more than 200,000 manually annotated sentences in FrameNet. A frame is a script-like schematic representation of a particular type of situation, object, or event with frame elements, which involving participants of sentence, props and other conceptual roles. A word with one of its senses is a lexical unit in a frame. FrameNet also includes relations between Frames. In this thesis, FrameNet interface from NLTK[7] is used.

SemiLink SemiLink[8] is a project which use mapping files to link different lexical resources together and link their annotated instances. In this thesis,

[3] https://nlp.stanford.edu/software/srparser.html

[4] http://verbs.colorado.edu/ mpalmer/projects/verbnet.html

[5] https://www.nltk.org/api/nltk.corpus.reader.verbnet.html

[6] https://framenet.icsi.berkeley.edu/fndrupal/

[7] https://www.nltk.org/api/nltk.corpus.reader.framenet.html

[8] http://verbs.colorado.edu/semlink/

mappings from FrameNet frames to VerbNet senses[9] and mappings from VerbNet roles to FrameNet arguments[10] are used.

WordNet WordNet[11] is a large lexical database of English. The words with a specific meaning are grouped as a 'synonym set', or 'synset', a list of synonymous. A word can have several meanings(senses) and can be identified in different synsets. Each synset comes with its definition and examples. The synsets are linked together with a IS-A relation in the WordNet Hierarchy. Hyponyms are children of a synset, and Hypernyms are parents of a synset. The similarity, i.e sementic relatedness, is quantified by comparing the depth of synsets in the Hierachy. Verb synsets towards the bottom express more particular action manners on an event. WordNet also contains a Lemmatizer, which can normalize a word for a given part of speech, when the word is in the dictionary of WordNet. In this thesis, WordNet interface from NLTK[12] is used.

3.3 GIS systems

A geographic information system (GIS) is a type of database operating and processing geographic data. Several GIS systems served as extra domain knowledge to identify the locations in the travel route.

GeoNames GeoNames[13] geographical database covers all countries and contains over 25,000,000 placenames of different languages, including their historical names. All features are categorized into one of the nine feature classes and further subcategorized into one of the 645 feature codes.

CHGIS CHGIS[14] stands for China Historical Geographic Information System. It is a database of populated places and historical administrative units for the period of Chinese history between 221 BCE and 1911 CE.

SRHGIS SRHGIS[15] stands for Silk Road Historical GIS. It is a open platform aiming to describe the Silk Road from Xi'an and the various geographic

[9] http://verbs.colorado.edu/verb-index/fn/vn-fn.xml

[10] http://verbs.colorado.edu/verb-index/fn/vn-fn-roles.xml

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

[12] https://www.nltk.org/api/nltk.corpus.reader.wordnet.html

[13] https://www.geonames.org/

[14] https://sites.fas.harvard.edu/ chgis/

[15] http://www.srhgis.com/dtcx

information along it, including the distribution of ethnic tribes, cultural heritage, the dynamics of deserts and oases, trade centres, commodities and types of trade, and other topics.

SRGIS SRGIS[16] stands for Silk Road GIS. It restores several trade routes of Silk Roads from Kashgar and Shache in China to Afghanistan and Tajikistan between the Wakhan Valley and Khansa in Kashmir. Through fieldwork, the team relocated all the important landmarks (mountain passes, settlements, ruins, etc.) in the region accurately.

[16] http://silkroad.fudan.edu.cn/project.html

Chapter 4

Corpus

In this chapter, the data set used in this thesis is introduced, and several analysis are made. Section 4.1 introduces the basic information about the historical data set, i.e. *The Travels of Marco Polo*. Section 4.2 describes the steps of how to make the data into a structured data set and how to make the gold standard annotations for evaluating and further usage. Section 4.3 describes how the gazetteer is built from the index of the books. Section 4.4 describes the quantitative and qualitative analysis of the data set and the characteristics saw in the analysis.

4.1 Overview

The Narrative, *The Travels of Marco Polo* was written in the 13th century by a Romance writer, Rustichello da Pisa, based on Marco Polo's dictations about his travels and experiences from Europe to Asia. The primary language of the book is Franco-Venetian, a literary language strongly influenced by the French language and widely used in northern Italy between the 13th and 15th centuries. The original title of the book is *Livres des Merveilles du Monde*. Although the original manuscripts were lost, the book was translated into different languages during Marco Polo's lifetime and various copies were made. From then on, the tales in the book influenced many travelers and adventurers in Europe, inspiring them to explore Asia to find the kingdoms and cities described in the book. Many of them tried to trace (part of) Marco Polo's route based on his description. The orientalists later rearranged the manuscripts, translated them into modern languages, and annotated the text according to other documents and the new records of explorers in 19th Century.

For this thesis, the English translation with the name *The Travels of Marco Polo* was used. It was translated and annotated by Henry Yule, later revised by Henri Cordier, based on several medieval manuscripts. The translators had kept the original writing of place names, personal names, and other named

Table 4.1: Overview of the Narrative

Part	**Description**	**Total Section**
Prologue	It describes briefly about the journey of the Polos'.	18
Book I	It describes the journey from the Lesser Armenia to the Court of the Great Kaan at Chandu.	61
Book II	It describes the Great Kaan, his capital city and the customs in Cathy and the journey through the Cathy and Manzi.	82
Book III	It describes Japan, the Archipelago, Southern India and the Coasts and Islands of the Indian Sea.	40
Book IV	It describes the wars among the Tartar Lords, and the Northern Countries.	34

entities, explained the objects to which they refer in the notes only, so that this work can still be considered as a historical document of the 13th century. The Book was published into 2 volumes. Both volumes are now transcribed into eBooks in the Project Gutenberg[1][2], which are mainly used in this thesis. The pages of scanned resources from New York Public Library are also used to check for typographical errors.

The Index of another translation from Hugh Murray was also utilized for constructing gazetteers in this thesis.

4.2 Data Preparation

4.2.1 Splitting

The Book consists of 2 parts:

The first part, a Prologue, is a brief narrative on how Nicolas and Maffeo Polo, the father and uncle of Marco Polo, travelled to Kublai Khan's court. It then describes the second journey with Marco Polo and their return to Persia through the Indian Seas. Due to its very different narrative focus and protagonist from the second part, this Prologue was not included in many editions.

The second part is the main portion of the work, later called the Context. According to the introductory notice at the beginning of the book, this part was traditionally divided into three books in Latin or early Italian editions. But in this edition, the last several chapters were separated to form Book IV. The descriptions of the books and their numbers of sections are shown in

[1] Volume 1: https://www.gutenberg.org/ebooks/10636

[2] Volume 2: https://www.gutenberg.org/ebooks/12410

Table 4.2: Example of a sentence and its annotations in Corpus

ID	**sentence**	**annotation**
1-1-1-0	There are two Hermenias, the Greater and the Less.	[['Hermenia', 'GPE']]

Table 4.1. Among all the Books, the Book II is divided into 3 parts. The first part is about Cublay Kaan , his capital city and the customs of his people. The second and the third part continued with the journey through Cathy and Manzi. For further analysis, the Context was divided into sections and then broken down into a sentence level granularity as shown in Table 4.2.

A unique ID was created to contain all the information about sections and positions of the sentences. The first Number indicates the Book index. The second Number indicates the Chapter index. The third Number indicates the position of the sentence in the chapter. The fourth Number records the Note index. With this format, all the sentences can be relocated easily in later tasks. Because the title of each chapter also contains Named Entities and carries necessary information about the travel route, so it is also stored in the corpus with the special sentence index 0.

The notes are seperated and stored in another dataframe table. They are only used as the assisted knowledge base to annotate the corpus and make the gazetteers. No analysis on the notes has been done in this thesis.

A dataframe table is used to store all the sentences with their corresponding information, and they are manually checked afterwards to make sure mistakes were not made in splitting.

4.2.2 Annotation

A reliable annotated corpus is the fundamental to evaluate the performance of the Nature Language Processing algorithms which are used to annotate the text. These standarded collections of annotations are called Gold Standard Annotations. The quality of the gold standard annotations impact the training of supervised-learning-based NLP algorithms directly. Although it takes long-time and big effort, usually these annotations should be created manually to guarantee the quality. In this thesis, several gold standard corpora have been made and can be used for further researches. The annotations of the corpora has been done semi-automatically, which means automatic algorithms were used at first and manual check were done later.

For the gold standard Named Entity Recognition annotations, the clear boundaries and the types of the named entities are important. Most of the observed named entities contain one or two tokens, still some contain more than three tokens. Firstly, the text was divided into sentence level. Then pre-

Table 4.3: Mapping between 18-Class NER Tags and 4-Class NER Tags

18-Class NER Tags	**4-Class NER Tags**
GPE	LOC
LOC	
FAC	
PERSON	PER
ORG	ORG
all other tags	MISC

trained Ontonotes (18-class) Named Entity Recognition model from Flair was used to identify the Named Entities in the sentence. In Table 4.3 a mapping between 18-class Tags to 4-class Tags is given. The 18-Class NER Tags can be easily turned into 4-Class NER Tags if needed. At last a manual check has been done to make gold standard annotations, by the assistance of elaborations from the index and the notes. The difference between the annotations from Flair model and the manual annotations is the performance of this model. It will be discussed later in Section 5.1.1.

The second gold standard annotations are about the motion verbs. With the help of the lexical resources, a list of possible motion verbs is built. These verbs are then marked in the sentences automatically. Inevitably, however, when these verbs are not interpreted as 'motion', they are also marked as motion verbs. Therefore is manual examination has been done to identify the real motion verbs.

The third gold standard annotations are about the motion events. With the utilization of the Part-of-Speech tag and parser, noun phrases and prepositional phrases belonging to the motion verb are identified. However, firstly, many of the identified noun phrases contain the common nouns (or phrases) which refer to the locations rather than the locative expressions themselves, while the place names they referred to are not in the same clauses or even in the same sentences. Secondly, for noun phrases that are not part of a prepositional phrase, their semantic roles are not easy to determine. Furthermore, a considerable number of phrases that do not contain a place are also marked just because they are noun phrases or prepositional phrases. Therefore, check of the identified phrases, determination of the actual locative expressions and their semantic roles requires manual inspection. The specific work will be described later in Section 5.1.2.

4.3 Gazetteer Preparation

Gazetteer is a set of lists containing named entities of various types, used to recognize the occurrence of the named entities in the context. The building of the gazetteer for this thesis based on the index from the back matter of *The Travels of Marco Polo*. Both indexes from Henry Yule and Hugh Murray are used in this thesis. Gazetteer helps to use the domain knowledge specific for the task.

Index is a list of words or phrases ('entries') with their related page numbers, so its contents can easily be looked up in the context. The format of index is showed in Example 4.3.1."Acomat Soldan" is the entry, "seizes throne of Tabriz" is a short elaboration of the entry, "ii" means the entry is in the second volume, "467" is the page number where the entry appears in volume 2, the other lines are the related information of the entry and the corresponding page number where they occur. In Project Gutenberg, all the page numbers in the index can be linked back to the page, so that it is easy for reader to look up. If a "n" is contained in the page number, like "470n" and "474n" in the last line of the example, means this entry appears in the notes. If a entry only exist in the notes, it will not be kept in the Gazetteer. Because these entries are meaningless for our research, for example the name of the modern adventurers whose records were used in the notes.

Example 4.3.1 *Acomat Soldan (Ahmad Sultan), seizes throne of Tabriz, ii. 467;*
goes to encounter Argon, 468;
rejects his remonstrance, 469;
defeats and takes him, 470;
hears of Argon's escape, is taken and put to death, 473;
notes on the history, 470n, 474n

The situation in Example 4.3.2 also exists in the index, which here the two entries are different named entities, but they shared the same first word. '——' represents the shared first word in the second entry.

Example 4.3.2 *Bitter bread, i. 110, 122*
—— water, i. 110, 122n, 194

Annotations The tags of the named entities, the same as the gold standard annotation of the content, use the Ontonote standard. The reason is that the writer believes that information under other tags, apart from the names of persons and locations, such as nationalities and products, is also valuable. Adding other tags in addition to the 4-class NER Tag alone, however, would affect the evaluation of the pre-trained model. Tagging named entities requires additional domain knowledge, so the job was done with a lot of online information and review of the original notes.

Ambiguity Resolution Due to the age of the book, the named entities are not well linked together in the index of this book. It is possible that the same named entity is recorded twice in the index under different entries, which cause ambiguities in the index. The alternative names of an entry can be given in following ways in the index:

1. The name of the entry is followed by the word in brackets. As in the underlined part of the Example 4.3.3, "Habsh" is the alternative name of the entry "Abash". If the bracketed word is also an entry in index, it does not mean that the relationship is bi-directional, i.e. the name outside the brackets does not always follow the other word in brackets.

2. Although a entry occurs in the text, no page number is given in the index for this entry, and a 'see' and another entry are given, at which point it is likely that the latter is the alternative name of the former. As shown in the bounded part of the Example 4.3.3, "Abyssinia" is the alternative name of "Abash". And "Abyssinia" is possibly a much well known name than "Abash" or "Habsh". But it may also be the case like in Example 4.3.4 that the latter entry is actually a superordinate concept to the former entry, rather than representing the same named entity.

Example 4.3.3 *Abash (Habsh), see* ***Abyssinia***

Example 4.3.4 *Animal Patterns, see Patterns*

3. The third is a bit more complicated, in that the two entries are not in each other's alternative name lists. But they have a same alternative name both in their lists, which undoubtedly means that the two entries refer to the same named entity. For example in Example 4.3.5, "Hang-chau fu" and "Cassay", they both have an alternative name "Kinsay", but they are neither in each other's alternative name lists or the alternative name list of "Kinsay". But all these three entries refer to the same named entity.

Example 4.3.5 *Cassay, see Kinsay*
...
Hang-chau fu, see Kinsay
...

Table 4.4: Example of a merged index

Entity Name	Elobrations	Related Contents	Alternative Name	Type
Kinsay	[’formerly Lin-ngan now Hang-chau fu’]	[’its surrender to Bayan;’, ’extreme public security;’ , ...	[’Capital’, ’ Khansa’, ’ Khinsa’, ’ Khingsai’, ’ Khanzai’, ’Cansay’, ’Campsay’, ’Kin-sai’, ’Quin-sai’, ’Kitvaal’, ’Quin-sal’]	GPE

Kinsay (King-szé, or “Capital,” Khansá, Khinsá, Khingsai, Khanzai, Cansay, Campsay), formerly Lin-ngan now Hang-chau fu, 11

There is a difference between the Murray version’s transcription of the manuscript and the Yule version. So to associate the entries from two indexes, in addition to considering all of the above satuations, it is necessary to compare similar entries back to the original text to determine if they are the same named entity. Here we give the Example 4.3.6, the entry refers to the same named entity as the entries in the Example 4.3.5.

Example 4.3.6 *Kitvaal or Quin-sal (Fang-tcheoufou), capital of Manji, 177,187.*

Conclusion Finally the gazetteer was merged and saved in a dataframe as the example in the Table 4.4. In addition to the Entity name, alternative names and Entity type, the elaborations and the related contents are also saved and merged from different entries if exist, for further usage.

Of all the entities in Gazetteer, more than 1200 named entities are generated. In Figure 4.1 compares the number of entries with various tags. The largest category is the "GPE" entities, with 335, along with 96 with "LOC" tags and 30 with "FAC" tags, while the number of entities with "PERSON" tags is 242. The entities with various ethnic or tribal are tagged with "NORP" tags, which is records 95 in the gazetteer. The "PRODUCT" entities, which are hard to be recognized in the context, reaches 245 in the gazetteer.

4.4 Data Analysis

4.4.1 Quantitative Data Analysis

The statistics in Table 4.5 on the corpus help to give insight into the size of the corpus and give an overview of the structure of the book. The Lexical Diversity

$$TTR = \frac{Vocabulary}{Token}$$

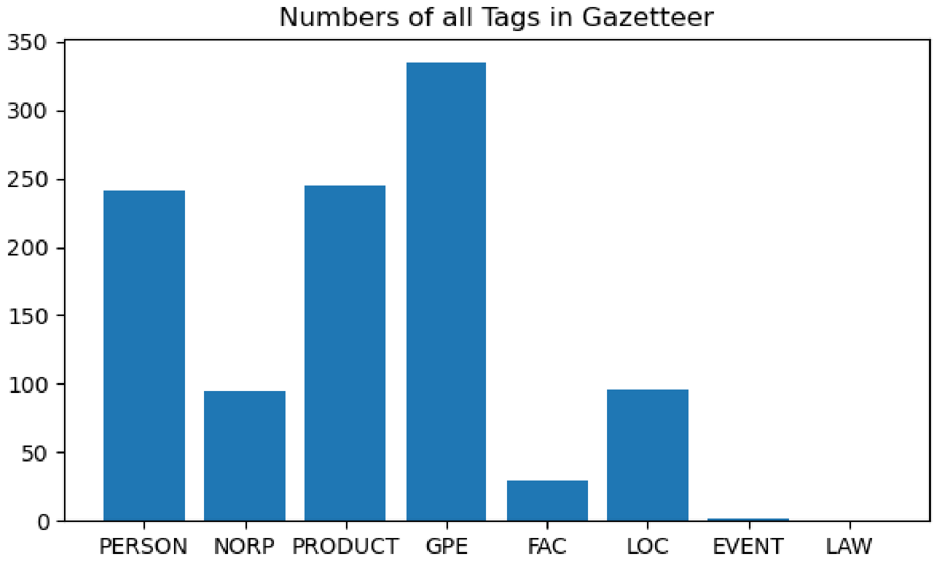

Figure 4.1: Numbers of different tags in Gazetteer

of the Context is 0.0667 and the of the Prologue is 0.1989, which implies that the information in the prologue is more general.

At the token level, the most common ngrams in the book were calculated after removing the stopwords and punctuations from the text, as showed in Figure 4.2. As showed in Figure 4.3, the most common bigram in the context of the book is 'great kaan', which is the most frequently used term for Kublai Khan in Marco Polo's travelogue. As showed in Figure 4.4, the other collocation means Kublai Khan is also very common as trigram. Locative expressions are also very common in trigrams, the trigrams "the city of", "the province of" and "the kingdom of" are all among the 20 most common trigrams.

When we counted the unigrams for each book, we found that it did suggest the content of each section. As in Figure 4.5, in Book I Marco Polo passes through many countries from Europe to Mongolia, hence the high occurrence of the word 'country'. Although 'province' and 'city' are frequently used in both Book I and Book II, 'city' appears more than 300 times in Book II as showed in Figure 4.6, far more than in any other books. This is because the journey in Book II passes through Cathy and Manzi, provinces with contiguous cities. In Figure 4.7, 'country' reappears in the top 20 most common words in Book III, along with 'kingdom', which has the similar meaning. Book IV, however, talks about the war between Tartar lords, so that the names of people, such as 'argon', 'caidu' and 'acomat', appear among the common words in Figure 4.8. Among the most common words in the Prologue, 'brothers' refers to Marco Polo's father and uncle.

Table 4.5: Descriptive information about corpus

	Sentences	**Tokens**	**Vocabulary**	**Average Tokens per Sentence**
Prologue	159	4937	982	31.0
Book I	1034	24519	3118	23.7
Book II	1662	45738	4351	27.6
Book III	949	23099	2859	24.3
Book IV	318	7699	1442	24.2
Total Context	3963	101055	6727	25.5

As can be seen in Figure 4.10, there are no uncommon verbs in the book that are used more often.

4.4.2 Qualitative Data Analysis

In this paragraph, we discuss the challenges we meet in extracting the travel route.

The narrative perspective of the context There are only very few sentences in which Marco Polo or his father or uncle are directly the subject, and most of the sentences describing the movements are in this way.

Example 4.4.1 *When you leave this city to travel further, you ride for seven days over great plains, finding harbour to receive you at three places only.*

The book does not devote all of its attention to describing Marco Polo's travels, instead it accounts a great amount of local folklore and historical legends which Marco Polo had heard of. Some locations nearby but not in the travel route are also included. It is difficult to tell whether a motion event is within Marco Polo's travel route simply by the sentences in Examples 4.4.1.

The order of the book The book was not written in strict chronological order, which creates some difficulties in tracing the travel route. For example:

Example 4.4.2 *Now you must know that between Anin and Caugigu, which we have left behind us, there is a distance of [25] days' journey; and from Caugigu to Bangala, the third province in our rear, is 30 days' journey.*

Examlple 4.4.2 indicates two routes, one between *Anin* and *Caugigu*, another between *Caugigu* to *Bangala*. So it should be the route between *Anin* and *Bangala* via *Caugigu*. But it is hard to identify, which of *Anin* and *Bangala* is the source and which of them is the destination.

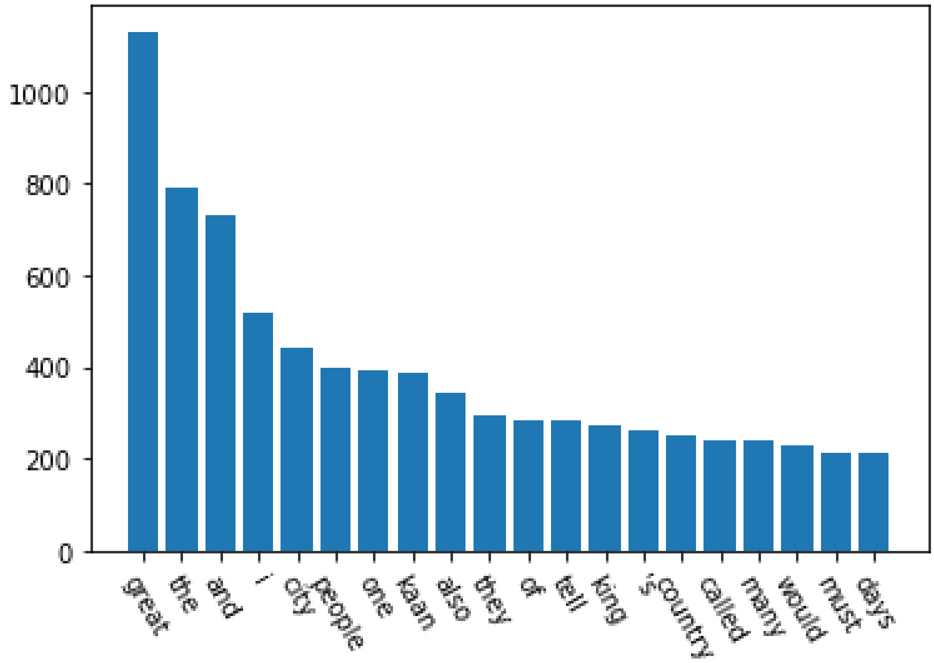

Figure 4.2: Most Common Unigrams

Ambiguity of location names The ambiguity exists in the locations names. The writer has observed several situations.

Location names can be changed due to hundreds of years. For example *Kinsay* now has the name *Hangzhou*, which is totally different. In this case, ancient location names are almost unrecognisable at present and require some extra knowledge to find out.

Ancient administrative areas would have changed, and it would be difficult to find a corresponding location or area on a modern map, even with the same or similar location name. As an example, the city of *Hormus* (Marco Polo's *Hormuz*) is now a city on an island, but the name used to refer to a city on the mainland near the sea.

Most importantly, Marco Polo did not master the local languages of all the areas he passed through. A estimate is that he mastered Persian and Tatar languages, by the Persian or Tatar name of locations he used instead of the name in local language. Therefore he also made many mistakes with location names.

The first result is, that he may identify one location with another due to the incorrect pronunciation.

Example 4.4.3 *"...Bangala is a Province towards the south, which up to the year 1290, when the aforesaid Messer Marco Polo was still at the Court of the*

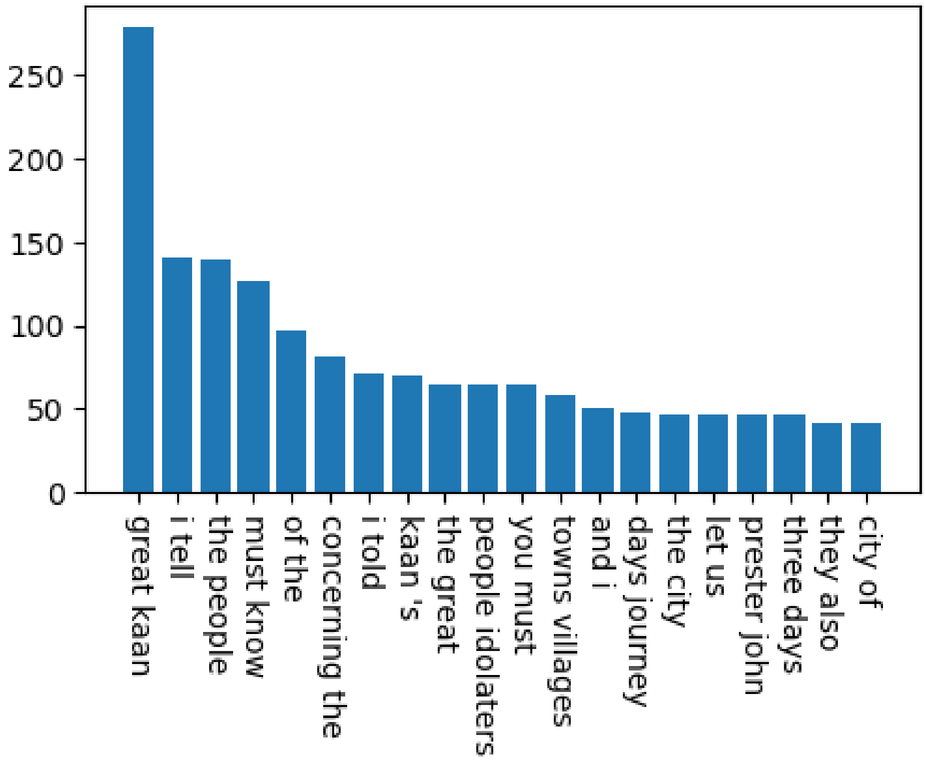

Figure 4.3: Most Common Bigrams

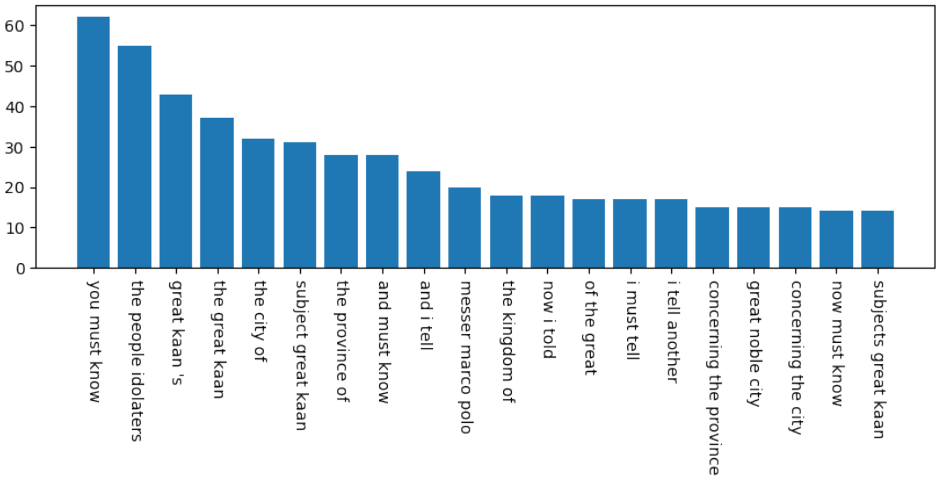

Figure 4.4: Most Common Trigrams

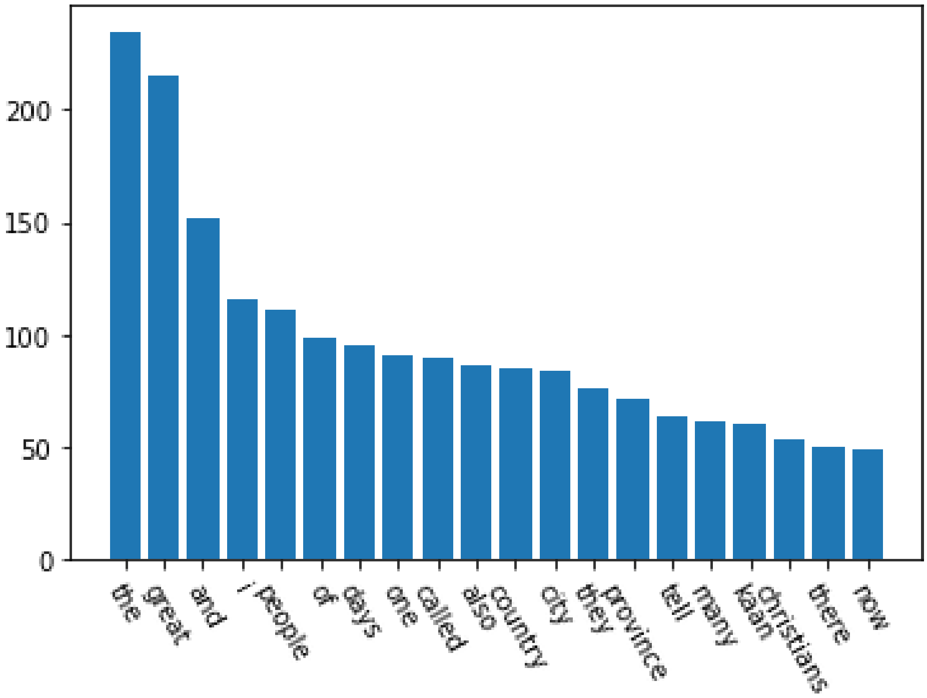

Figure 4.5: Most Common Unigrams in Book I

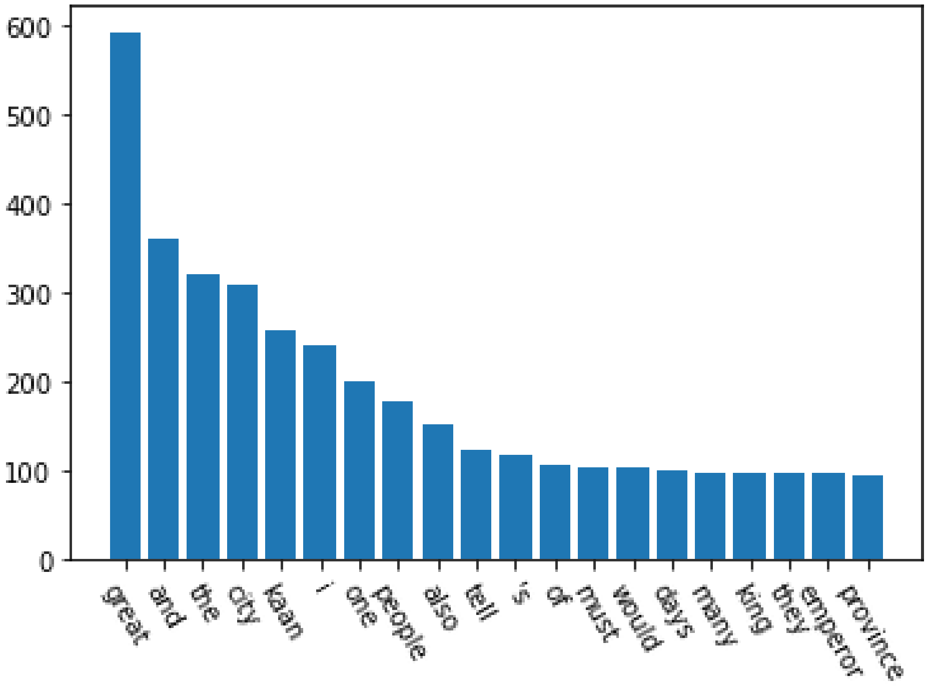

Figure 4.6: Most Common Unigrams in Book II

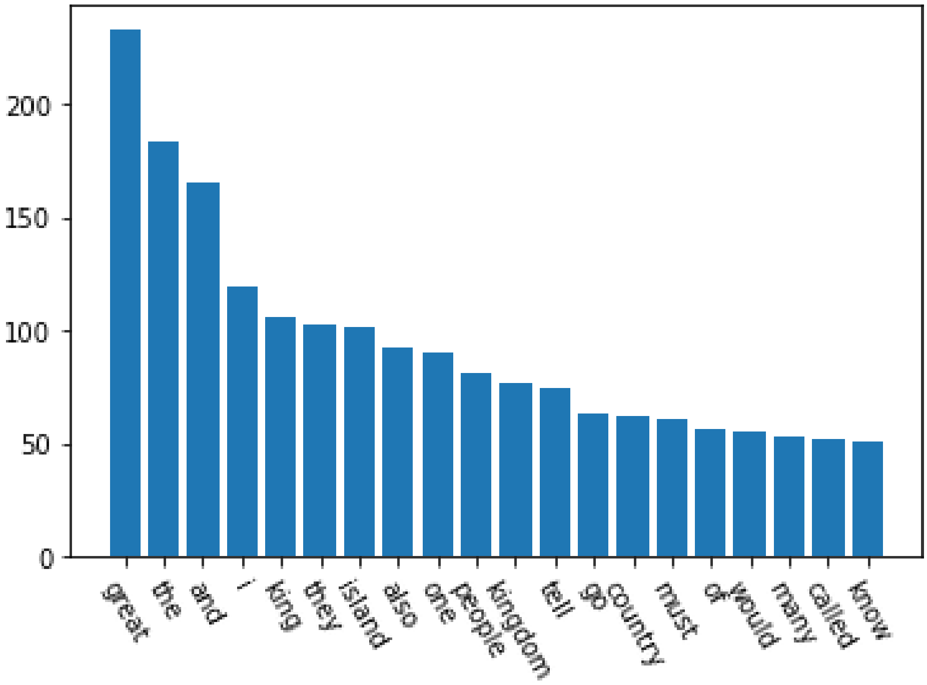

Figure 4.7: Most Common Unigrams in Book III

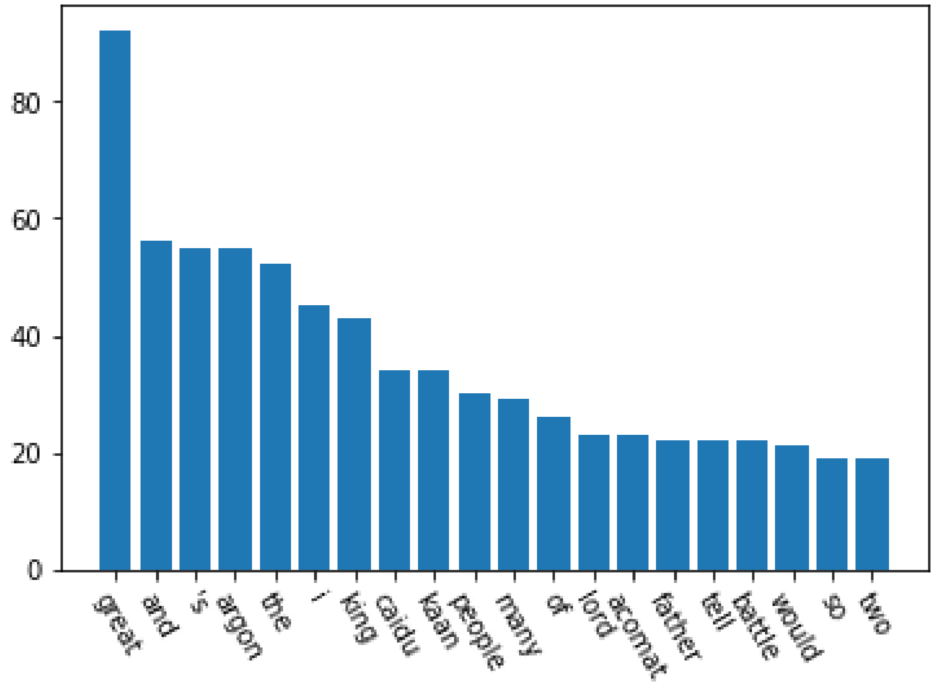

Figure 4.8: Most Common Unigrams in Book IV

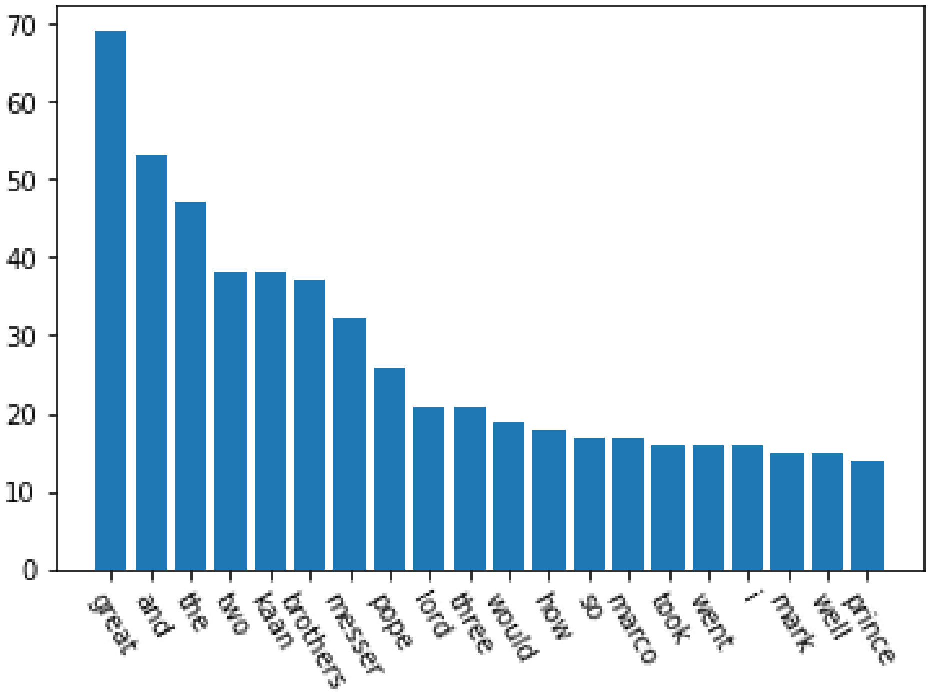

Figure 4.9: Most Common Unigrams in Prologue

Great Kaan, had not yet been conquered..."

As in Example 4.4.3, *Bangala* is the ancient name of *Mangala*, which was never conquered by *Kublai Khan*. The most possible explanation is that he made a mistake with *Pagu*, a city in Mien and near his travel route in Southeastern Asia, also pronounced similar to *Bangala*.

Another mistake which Marco Polo made very often, is the the locations, of which names end with with 'chu', which are usually cities, and those end with 'fu', which are usually bigger. But since the pronunciations are similar to him, he simply mixed the two characters up. For example, *Sindachu* by him was in fact a 'fu' at his age.

The last of these comes from the fact that Marco Polo did not choose the most accurate location names. Some of the location names derived from the names for certain landscapes in local language. For example, he passed through *Pamir* on his journey, although we now also use the name Pamir for the big mountain range between Central Asia, South Asia and East Asia. But the word ‘Pamir’ means ‘a flat plateau or U-shaped valley surrounded by mountains‘[3] in local language. There are, altogether, eight Pamirs plateaus in

[3] https://en.wikipedia.org/wiki/Pamir_Mountains

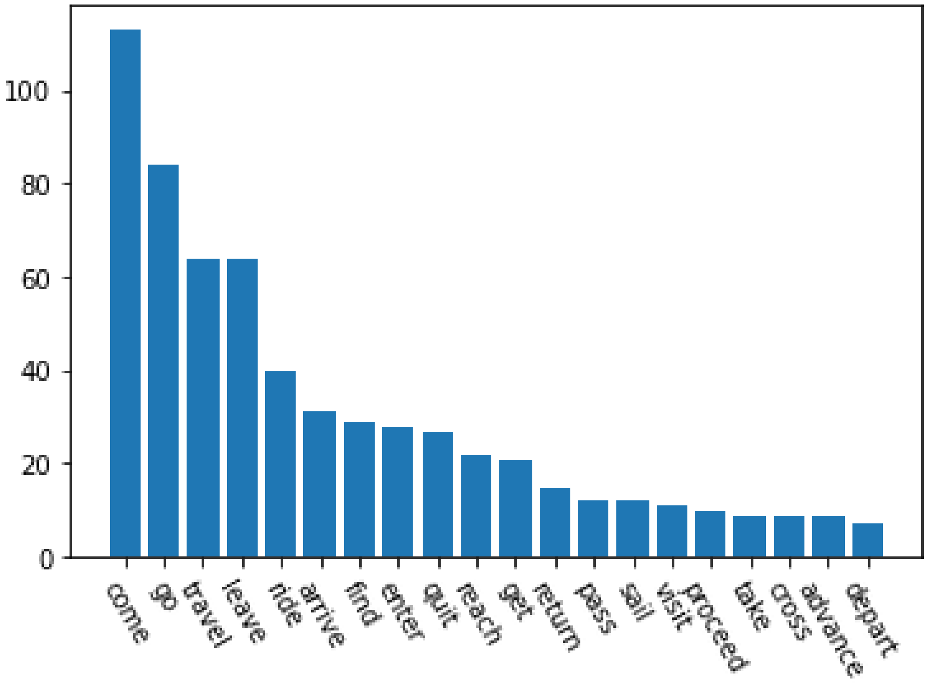

Figure 4.10: Most Common Motion Verbs in Context

the area, with very similar landscapes. It has been a long debate on which of them is the exact Pamir passed by Marco Polo, in the notes of this book, at least three opinions are gived on this issue.

Chapter 5

Experiments

5.1 Methods

5.1.1 Location Entities Identification

In order to identify Marco Polo's travel route in the text, the location entities he mentioned in the Context had to be identified firstly, and we used named entity identification to approach this goal. The task of the NER can be divided into two parts, the first is to determine the boundaries of named entities, and the other is to determine their categories. The categories we are most interested in are the 'GPE', 'LOC' and 'FAC' tags in Context, while the other tags can be used in further studies. This paper uses three methods to approach this goal:

1. A preliminary named entity recognition was done by using the pre-trained large 18-class NER model for English that ships with Flair on unlabelled text, in order to create gold standard annotations on this basis manually. Flair uses Wordembbings and achieves good results on the contemporary text. An f1 score of 94.36 was obtained on CoNLL-03 Corpus and 90.93 on Ontonote. The results of using the pre-trained model are also recorded in this paper and later compared with the gold standard annotations. It represents the performance of the model, which was trained based only on the contemporary text, on the historical text.

2. After completing the gold standard annotation, the authors fine-tuned the pre-trained model with the method of supervised learning, which means a function using labeled training data. Flair's label, i.e the tag dictionary, uses an inside-outside-beginning format to indicate boundary and type. the Table 5.1 shows an example on the inside-outside-beginning format, with '-I' for inside the tag and 'O' for outside the tag, and '-B' for beginning of the tag. Fine-tuning is run using an LSTM to learn in both directions. The training is

Table 5.1: Example of Inside-Outside-Beginning format NER

In	O
GEORGIANIA	I-LOC
there	O
is	O
a	O
King	O
called	O
David	B-PERSON
Melic	I-PERSON

based on small and imbalanced corpus.

3. After examining the length of the named entities in the gazetteer, the writer find that the longest named entity is of 5 words. So the n-grams of the sentences are obtained from n=5 to n=1 to look up in the gazetteer. The match lists are compared with the gold standard annotation to evaluate the performance. Two gazetteers are used in the paper, the first one is a gazetteer built from index and the test set also comes from context for performance evaluation. The second gazetteer is built from the gold standard corpus of the Context, which is totally correct for the context, containing only annotations from the context. This Gazetteer is used on Prologue to evaluate the performance of using gazetteer on similar text.

5.1.2 Motion Events Extraction

Events extraction is an information extraction task, trying to identify the fundamental participants, e.g 'Who did What to Whom and Where and When' in the event. The core of an event is the verb. In our task, we are most interested in the verbs represent motions. We divide the process of finding motion verbs into three steps:

Generating motion verb list from lexical resources;

Marking the verbs in the sentences, and finding the base forms of the verbs;

Comparing the base forms of the verbs to the motion verb list to identify the motion verbs.

We use both VerbNet and FrameNet to generate the motion verb list. As introduced in Section 3.2, in the hierarchy of VerbNet, the verbs classes under a top-level number share a similar semantic meaning. Class number 51 represents 'Verbs of Motion'. By checking the semantics of the verb classes, we choose "'reach-51.8', 'meander-47.7', 'escape-51.1-2-1', 'escape-51.1-1', 'escape-51.1-2', 'leave-51.2-1', 'roll-51.3.1', 'run-51.3.2', 'nonvehicle-51.4.2'" as the subclass

Table 5.2: Verbs in Motion Verb List

	Before WordNet Similarity	**After WordNet Similarity**
VerbNet	184	147
FrameNet	445	299

list. All the lemmas, i.e words, are put in the motion verb list. The important semantic roles of motion verbs from VerbNet are 'Initial_Location', 'Destination', 'Location' and 'Source'. By mapping from the Verb-Net roles to FrameNet argument, 32 frames containing these frame elements are found. After checking the definition of frames from FrameNet, 23 of them are considered to be possible motion verb frames. All the lexical units, i.e words, in the frames will be put in the motion verb list. Later maximal Wu-Palmer Similarity [22] of every verbs in the motion verb list with the seed verbs, 'go','travel','leave','arrive','reach', is calculated. It returns a score based on the position of the verb and seed verbs in WebNet Hirachy, describing how similar the verb is with the seed verb by calculating the depth of the two verbs in the taxonomy and that of their Least Common Subsumer.

The authors then tagged the sentences with POS Tag. It can be observed that a motion event does not always occur as the main predicate of the sentence, it can also be other participants of sentence, as in the Example 5.1.1. Therefore, all forms of verbs need to be checked, i.e. all words in the sentence that have 'VB' in the Tag. The motion verb list obtained from FrameNet and VerbNet contains only the base forms of the verb. So in order to compare the marked verbs in sentence with the motion verb list, the base forms of the verbs are generated by WordNet Lemmatizer.

Example 5.1.1 *"On leaving the Palace of Mangalai..."*

5.1.3 Route Extraction

In a sentence, location is contained within the noun phrase as the object of the verb. The transitive verb directly governs the noun phrase, and the intransitive verb governs the noun phrase by governing the prepositional phrase. In one sentence or clause, it it possible that more motion verbs occur. In this step, therefore, the author re-divides the sentences, which contain motion verbs into shorter sentences or phrases, according to the motion verb phrase or comma.

In the shorter sentences or phrases, a parser is applied to generate the syntactic tree. A verb and the phrase it governs will form a verb phrase. The writer therefore marks prepositional and noun phrases in verb phrases, where the subject verb is a motion verb, and looks for locations within them. These locations, which may be the starting points, the end points or the locations where the motion events happened, and these semantic roles are confirmed manually. Sometimes the phrases contain indicative pronouns referring to a previously mentioned location, which is sometimes not even in the chapter, needs to be looked up based on human understanding, as showed in Example 5.1.2.

Example 5.1.2 *"...and proceeds to that city which he has built..."*

The modern names of the locations are confirmed based on the extra knowledge from the notes of the book and GIS database.

In addition, prepositions expressing direction are marked out as the 'Location' role of the motion verb. They help to confirm whether this is indeed the direction between the former extracted location and the latter extracted location.

As stated in the observation, due to the narrative perspective of the context, it is not possible in this work to determine whether locations are included in Polo's travel route by the agent of the motion event. It can only be confirmed by reference to the direction between locations.

5.2 Results

The performance of the algorithm is evaluated by calculating the result of the algorithm and the error, comparing with respect to the gold standard dataset. Instead of observing only the correct portion simply, the confusion matrix is used in this thesis. In confusion matrix, the result of the algorithms is divided into four sets: True positive (TP) set, in which the objects are relevant and marked with right tags. False positive (FP) set, in which the objects are not relevant and marked with correct tags. False negative (FN) set, in which the objects are relevant but not marked with the correct tags. True negative (TN) set, in which the objects are not relevant and not marked.

From these four sets, we can get precision

$$Precision = \frac{TP}{TP + FP}$$

and recall

$$Recall = \frac{TP}{TP + FN}$$

Precision shows the portion of items we marked are actually right. Recall shows the portion of relevant items are marked.

F-measure

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

is the harmonic mean of precision and recall, combines them to give a single score of performance.

For a multi-tag task, for example Named Entity Recognition, confusion matrix and the corresponding precision, recall and F-measure can be calculated for every tag. In this thesis, micro F1 is used to calculate the total f1 score, which means the TP, FP, FN of the whole result are used for the total f1 score.

Location Entities Identification Figure 5.1 shows the performance of three algorithms on Location Tags ('GPE', 'LOC', 'FAC'), in which we are most interested, we see that the fine-tuned model achieves a worse F1-score than pre-trained model. It can mark more relevant locations than pre-trained model, but not all the entities it marks are locations. The reason is that the imbalance number of tags disrupt the identifying of different tags. By looking at Figure 5.2 and Figure 5.2, the fine-tuned model improve the performance on the tags which have very poor performance before fine-tuning. Concerning its high precision, we estimate that after fine-tuning, it learns some features of those tags. As a result, the average performance of the fine-tuned model on all tags is much more better than pre-trained one.

The Gazetteer reaches a very high precision, recall as well as the F1-score on location identification, which means it is, at least for our task, is the most powerful. But it does not performance well on all tags. The reason why is that a few 'PRODUCT's are not treated as named entites but as normal nouns in the corpus. On the other hand, the nouns can not be good normalized either. Because the Lemmatizer works only when the word is in its dictionary, but some of the words are not modern English, which still raise the problem because of the change of English language.

In Figure 5.5, we compares the performance of Gazetteer on context and the gold standard gazetteer of context on prologue. We see that the average performance does not decrease. But when items does not exist in gazetteer, it can not identify, which causes the decreasing of performances on Location Tags.

Motion Events Extraction As showed in Figure 5.6, almost all the motion verbs are marked by our algorithms, but only a small part of marked verbs are real motion verbs. Motion verbs contain some very commonly-used verbs like 'go'. And according to our algorithms, when we mark a verb in the motion verb list, we do not check its sense.

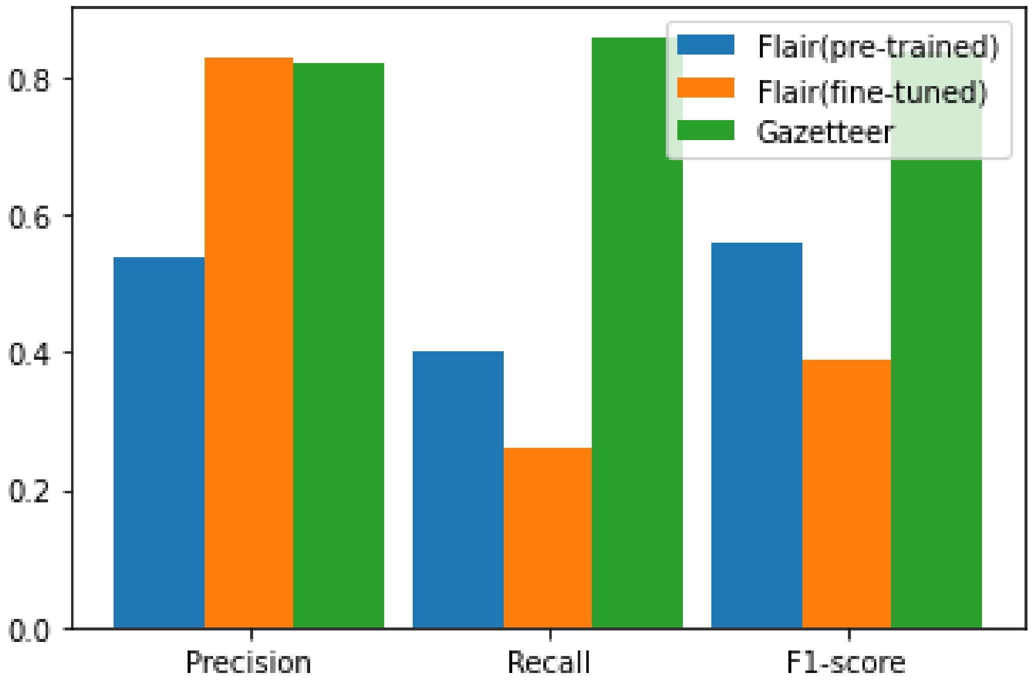

Figure 5.1: Comparison of three Algorithms for Location Tags

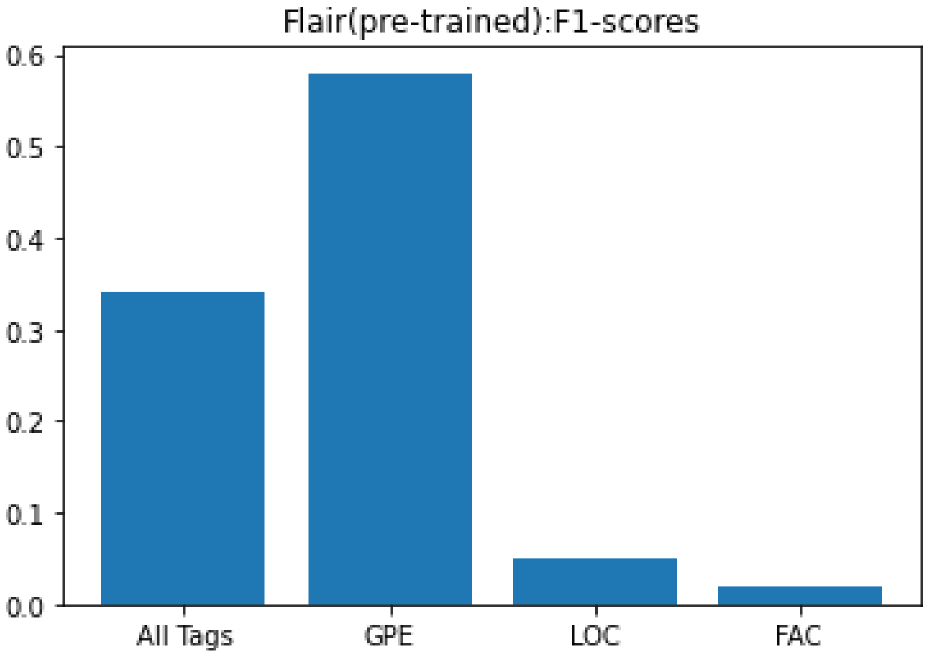

Figure 5.2: Flair Pre-trained Model: Total F1 score and F1 of different Location Tags

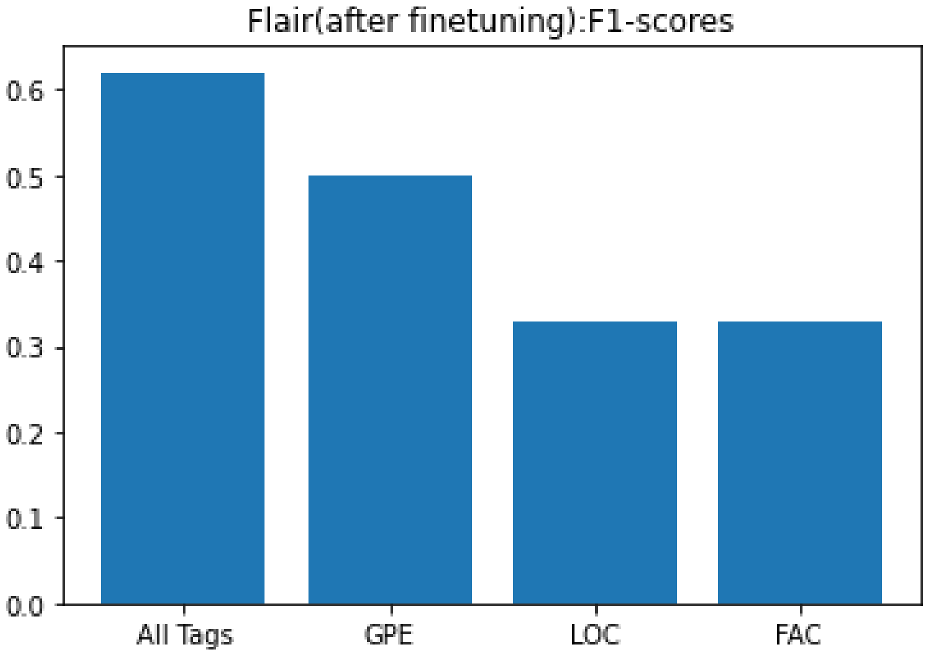

Figure 5.3: Flair Model after Fine-tuning: Total F1 score and F1 of different Location Tags

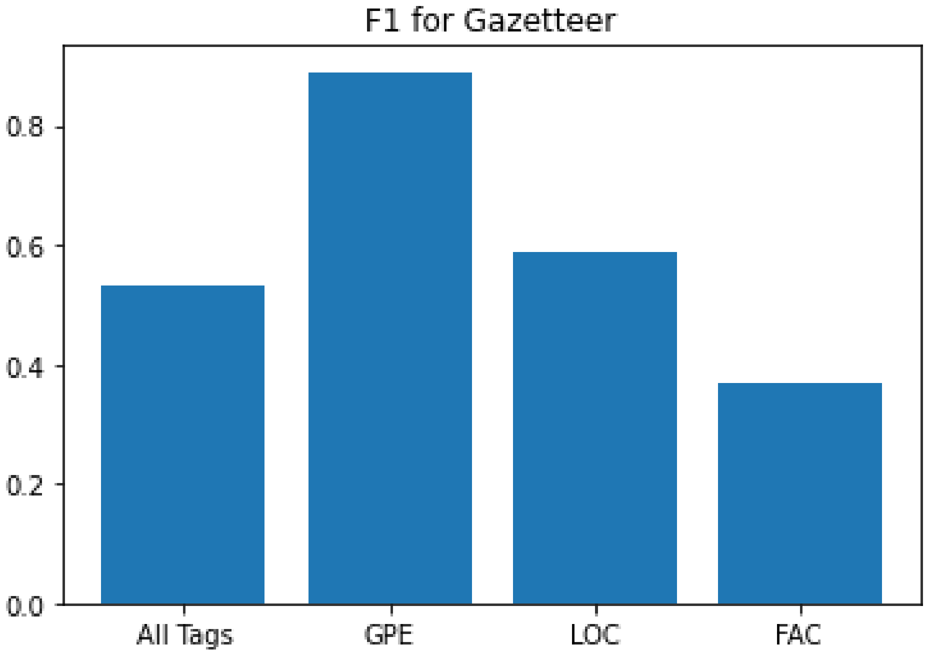

Figure 5.4: Gazetteer: Total F1 score and F1 of different Location Tags

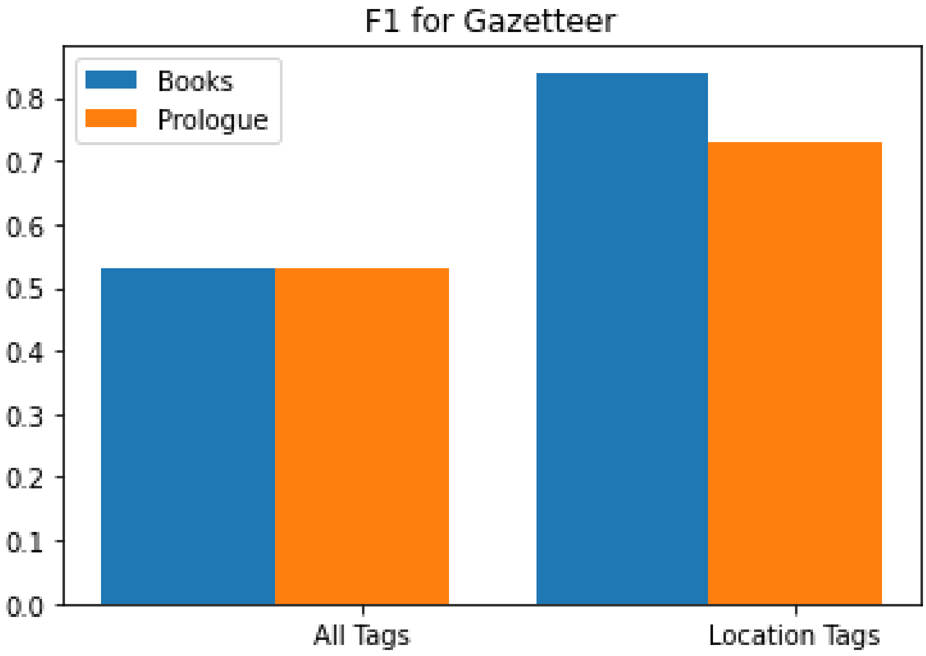

Figure 5.5: Gazetteer: Performance Comparison in Context and Prologue

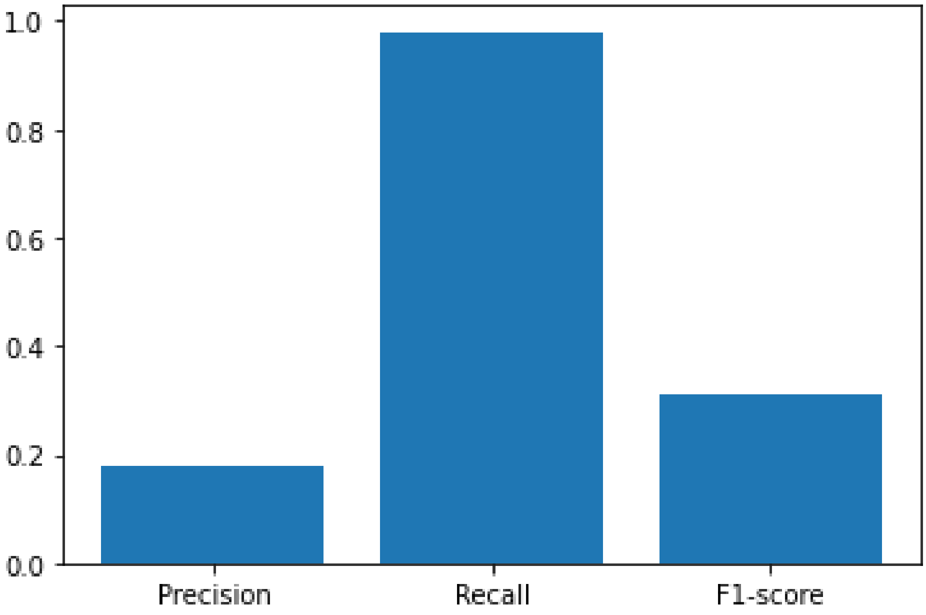

Figure 5.6: Motion Verb Detection

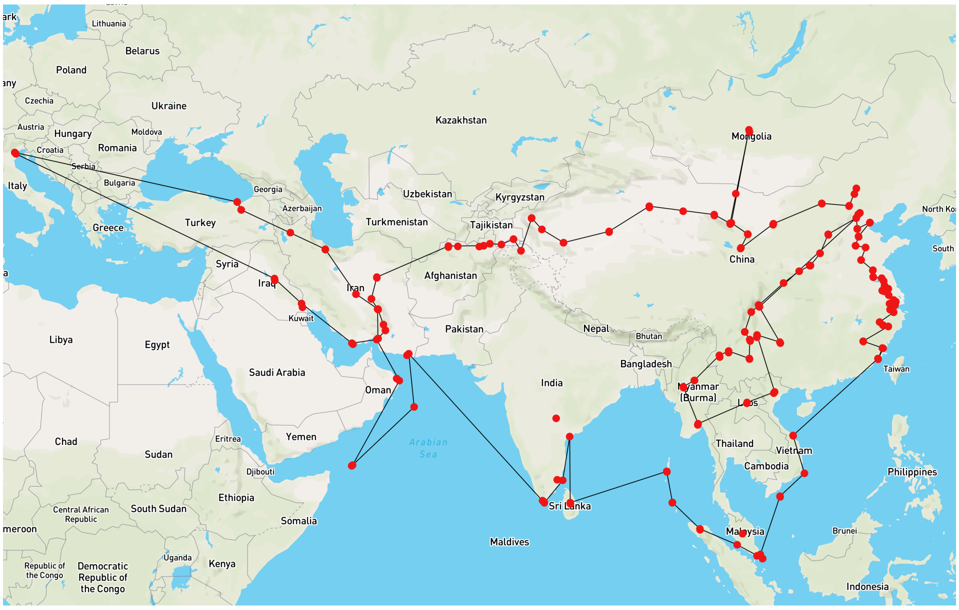

Figure 5.7: Travel Path of Marco Polo

Route Extraction After route extraction, 398 short sentences, which express motion of Marco Polo and contain relative exact sources, destinations or locations, are found. The writer put 133 locations on the Map 5.7 and link them with possible chronological order. The Result is saved as a kml file.

Chapter 6

Conclusions

6.1 Conclusions

In this paper, through the extraction of Marco Polo's travel routes from his travelogue, which preserves the 13th century proper names, named entity recognition, thematic (movement) verb extraction and route reconstruction are implemented. Their results are evaluated and the difficulties experienced are recorded.

In the area of named entities recognition in historical texts, the high accuracy of gazetteer based on special domain knowledge remains irreplaceable by machine-learning-based models. But building a gazetteer requires a great amount of manual labor and a significant amount of specific knowledge. The paradox of building a gazetteer is that the information extracted from historical documents may not be meaningful to the expert if he can build the gazetteer with his own knowledge; if the expert's knowledge is not sufficient to cover the historical texts, then a gazetteer based on his knowledge may not extract enough useful information from the historical documents.

In addition, although the machine-learning-based models cannot replace gazetteer for the time being, they can learn the features of the tags after fine-tune and the accuracy (f1-score) is improved. The low accuracy of the models on historical documents is probably due to the fact that their training material is mainly contemporary texts. If enough historical sources can be used for training, accuracy is sure to improve considerably. Furthermore, the size of the corpus, the number and balance of tags, and the quality of the gold standard annotations, also affect the accuracy of the model.

The extraction of thematically specific verbs is still limited by the ambiguity of the verb's lexical meaning. With the categorization of lexical resources, a list of synonymous or near-synonymous verbs can be obtained, which basically covers the verbs to be extracted. However, the problem is that the verbs in the list are also marked when they are using other senses. Therefore, an algorithm needs to be designed to disambiguation the extracted verbs.

In the present work, the main problem with the extraction route lies in the indicator pronoun. The author observed that the common practice of selecting location entities in close proximity does not perform well in texts that are not ordered exactly chronologically. As a consequence of manual processing, the use of proximity-selected location entities needs to be checked with additional information. In this thesis, this means the actual location of the location entity, the direction and distance between the two locations. Extracting information based on a pattern is valid to a certain extent, but confirming the roles of entities for forms that do not fit the pattern still requires more manual checking.

6.2 Future Work

This thesis has made a relatively deep mining of the information in *The Travels of Marco Polo*. But this mining is mainly focused on the single books of *The Travels of Marco Polo*. If we consider the traditional ways of historians in verifying a particular name, comparisons between the different collections of travel literature, e.g *The Rihla by Ibn Battuta*, from the same period or even different period can be made. Due to the complexity of natural language, this job is currently still done manually, but one could consider using geographical features, place name pronunciation, etc. as features to extract cross-textual information about the same named entity and make comparisons with the natural language processing techniques.

In addition to extracting information on location names, the gold standard annotations created in this thesis with the Ontonote standard also extracts information such as ethnicity and religious distribution. Marco Polo's travels from Europe to East Asia, passed the Middle East and the Pamir Plateau, can be used as additional material for the study of these regions. Based on this fact, it is hoped that a knowledge base in this field can be built up in the future.

As a long debated theme, whether Marco Polo visited East Asia or not can be verified by comparing the accounts in *The Travels of Marco Polo* with the real world. In the travelogues, although in this thesis it is not made, a systematic comparison, the locations on his route are very similar to the actual locations. For example, his descriptions of the landscape of the *Pamir Plateau* (Great Pamir), of the stone bridge over the *Pulisanghin*, and of the *West Lake* in *Kinsay* (Hangzhou) are all very close to the real world. If a systematic comparison could be made between more real locations and Marco

Polo's accounts, the accuracy of his accounts could be calculated. This can answer the question to a large extent of whether Marco Polo visited East Asia or not.

When evaluating the performance of the gazetteer, we used the gold standard gazetteer obtained from the Context to understand the named entities of the prologue, and achieved relative good results. As mentioned in chapter 4.2, the narrative of prologue is very different from the main text. The results achieved by this demonstrate that a specific gazetteer can achieve relatively good results in similar historical texts. Then, turning more historical documents into structured data, building Corpora of historical documents and training models based on them will greatly help the accuracy of automated information extraction from historical documents. This can be a research direction in the future.

Appendix A

Ontonotes Tags

Table A.1: Ontonotes Tags

Tag	**Meaning**
CARDINAL	cardinal value
DATE	date value
EVENT	event name
FAC	building name
GPE	geo-political entity
LANGUAGE	language name
LAW	law name
LOC	location name
MONEY	money name
NORP	affiliation
ORDINAL	ordinal value
ORG	organization name
PERCENT	percent value
PERSON	person name
PRODUCT	product name
QUANTITY	quantity value
TIME	time value
WORK_OF_ART	name of work of art

Appendix B

Part-of-Speech Tags Word Level

Tag	**Description**
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner
WP	Wh-pronoun
WP$	Possessive wh-pronoun

Appendix C

Part-of-Speech Tags Phrase Level

Tag	**Description**
ADJP	Adjective Phrase.
ADVP	Adverb Phrase.
CONJP	Conjunction Phrase.
FRAG	Fragment.
INTJ	Interjection.
LST	List marker.
NAC	Not a Constituent.
NP	Noun Phrase.
NX	Used within certain complex NPs to mark the head of the NP.
PP	Prepositional Phrase.
PRN	Parenthetical.
PRT	Particle.
QP	Quantifier Phrase ; used within NP.
RRC	Reduced Relative Clause.
UCP	Unlike Coordinated Phrase.
VP	Vereb Phrase.
WHADJP	Wh-adjective Phrase.
WHAVP	Wh-adverb Phrase.
WHNP	Wh-noun Phrase.
WHPP	Wh-prepositional Phrase.
X	Unknown, uncertain, or unbracketable.

Bibliography

[1] Dirk Ahlers. Assessment of the accuracy of geonames gazetteer data. In *Proceedings of the 7th workshop on geographic information retrieval*, pages 74–81, 2013.

[2] Adrien Barbaresi. A constellation and a rhizome : Two studies on toponyms in literary texts. In *Proceedings of ACL/IJCNLP 2021: System Demonstrations*, 2018.

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

[4] Davide Buscaldi. Approaches to disambiguating toponyms. *Sigspatial Special*, 3(2):16–19, 2011.

[5] Catherine Emma Jones, Marta Severo, and Daniele Guido. Socio-spatial visualisations of cultural routes. *Netcom*, 2018.

[6] Anna Korhonen and Ted Briscoe. Extended lexical-semantic classification of english verbs. In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL 2004*, pages 38–45, 2004.

[7] Ruth Mostern and Elana Gainor. Traveling the silk road on a virtual globe: Pedagogy, technology and evaluation for spatial history. *Digit. Humanit. Q.*, 7, 2013.

[8] Patricia Murrieta-Flores, Christopher Donaldson, and Ian Gregory. Gis and literary history: Advancing digital humanities research through the spatial analysis of historical travel writing and topographical literature. *Digital Humanities Quarterly*, 2017.

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

[10] Bruno Pouliquen, Marco Kimler, Ralf Steinberger, Camelia Ignat, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger,

Ann-Charlotte Forslund, and Clive Best. Geocoding multilingual texts: Recognition, disambiguation and visualisation. *ArXiv*, abs/cs/0609065, 2006.

[11] Martin Riedl and Sebastian Padó. A named entity recognition shootout for German. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 120–125, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[12] Ellen Riloff. Automatically constructing a dictionary for information extraction tasks. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, page 811–816, 1993.

[13] Kirk Roberts, Cosmin Adrian Bejan, and Sanda Harabagiu. Toponym disambiguation using events. In *Twenty-Third International FLAIRS Conference*, 2010.

[14] Stefan Schweter and Alan Akbik. Flert: Document-level features for named entity recognition, 2020.

[15] Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, S. Kulick, Neville Ryant, and Xiaoyi Ma. From light to rich ere: Annotation of entities, relations, and events. In *EVENTS@HLP-NAACL*, 2015.

[16] Amelia Carolina Sparavigna. From sheberghan to kashgar in the travels of marco polo. *Philica*, 08 2017.

[17] Amelia Carolina Sparavigna. The road to xanadu in the travels of marco polo. *Philica*, 2017, 08 2017.

[18] Amelia Carolina Sparavigna. From Kashgar to Xanadu in the Travels of Marco Polo. working paper or preprint, April 2020.

[19] Amelia Carolina Sparavigna. Marco Polo in Persia. working paper or preprint, August 2020.

[20] 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.

[21] 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, 2018.

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

[23] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November 2020. Association for Computational Linguistics.

[24] Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*, 2020.