English Readings of Chinese Characters in Japanese Names

Lars Holdaas

Supervisors:
Rune Sætre
Björn Gambäck

Norwegian University of Science and Technology
Department of Computer and Information Science

Abstract

Finding the correct reading of a Japanese name spelled with Chinese characters (Kanji) is a recurring problem for foreigners and Japanese nationals alike. The Japanese reading of a Chinese character varies with the region and the time period.

Although dictionaries exist that contain the readings of all Chinese characters used in Japan, this Master thesis focuses on building a database system that contains not only Chinese characters and readings, but also contextual data specifying which region and during which time periods certain characters were read in certain ways. This database will be helpful for linguists and name researchers that aim to track both readings of Chinese characters and which names were used throughout Japanese history.

In addition, the Master thesis uses the Japanese Wikipedia as input to create the data necessary to populate the database. As such, a large part of the thesis focuses on using simple natural language processing methods to extract certain information from the Japanese Wikipedia, and will function as a proof of concept that such a database can be created using publicly available information.

Contents

TODO: Liste med forklaringer av enkelte japanske ord som går igjen (Kanji, Onyomi, Kunyomi). Kanskje også ord som forklarer NLP-uttrykk? Tokens, Tokenizer etc? Hør med Rune/Björn.

1 Introduction

This chapter states the background and motivation, the thesis research questions, the procedure of achieving states goals and thesis structure.

1.1 Motivation

Readings of Japanese names that use kanji is a reoccuring challenge both for Japanese and non-Japanese alike. Although existing dictionaries already exist, there is no contextual database that allows users to not only find candidate readings of kanji names, but with context such as time period and geographical location further find the likelyhood of which reading should be used. The thesis therefore focuses on having such a database as the final product of the research.

Wikipedia has in the recent years become more of a focus in various research, due to the vast data available and the relative uniformity of how it is presented. Although recent research often focuses on using Wikipedia with relatively complicated natural language processing technique [1], this thesis will focus on simpler information extraction methods due to the scope of the thesis also focusing on a larger system.

Finally, the finished result will hopefully be useful for the following:

- Linguists that wants to trace pronounciations of common kanjis through various time periods and locations
- Name researchers that focus on where various names are used
- Computer scientists that are interested in using Wikipedia in similar way to construct knowledge systems
- General users that want to look up different names and get an understand of the origins of certain readings

1.2 Research Questions

- In which ways is the Japanese Wikipedia suitable as the basis for building knowledge databases such as the one proposed?
- To which degree are relatively simple natural language processing techniques sufficient for extracting the data required for such a database?
- Will there be sufficient correlation between readings of names and time periods/provinces that such a database can be used for further research?

1.3 Objective

The objectives of the thesis are the following:

1. Create a set of natural language processing algorithms that can sufficiently separate articles on the Japanese Wikipedia that are on persons with kanji in their names from articles that are not

2. From these articles create a database of names written both in kanji and the Roman alphabet as well as information about the time period and geographical location the name was used

3. From this database create a searchable web system where users can easily find this information for further use in their research

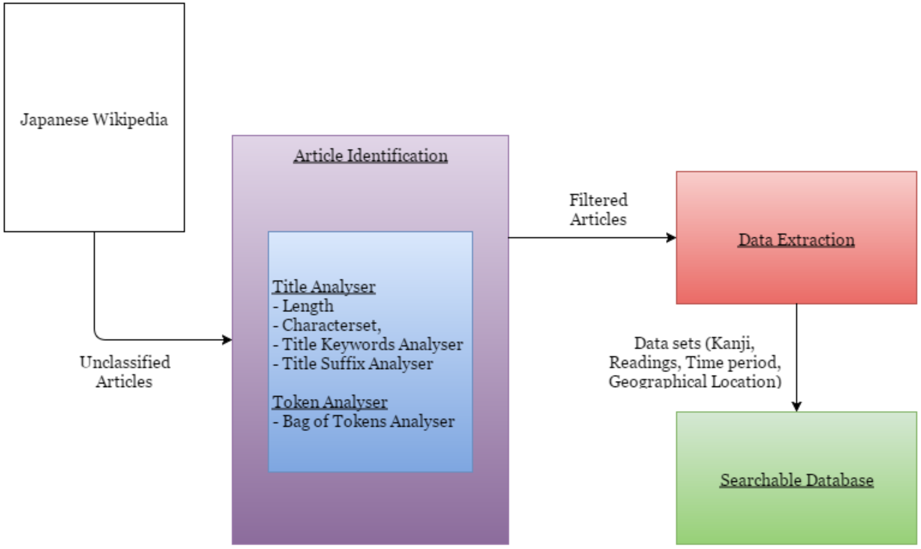
1.4 Research Approach

- Prototyping the suggested architecture and software to prove the concept

- Test article analyser hit-rate and compare to manually prepared samples

- Use results from the manual testing as basis for machine learning

The complete program will be structured in the following way:



1.5 Thesis Structure

The thesis uses the following structure:

1. Introduction: Defines the task and the planned approach to achieve the objectives

2. Theoretical Background: Defines the theoretical background of the following parts, including explaining in sufficient detail the problem of kanji readings to those unfamiliar with the Japanese language

3. Article Identification: Explains the implementation of the first part of the software which should identify which articles are relevant and which are not

4. Data Extraction: Explains the implementation of the data extraction software that should find information from the articles and put it into a format easily interpreted by a computer

5. Searchable Database: Explains the implementation of the searchable systems where users can easily find the data provided by the system constructed in this thesis

2 Theoretical Background

2.1 Kanji

漢字

A Kanji is a single non-phonetic symbol meant to represent a concept. Comparing Kanjis to western languages, they can be compared to the use of Arabic numbers to represent numbers. Arabic numbers can be used as a symbolic representation of a value between individuals who do not necessary share a common language, and who would be unable to communicate the value using words alone. Arabic numbers contain no information whatsoever as to how to pronounce the values they represent, but are extremely efficient in its way to convey the concept of value (compare reading ”3204+1381” to ”three-thousand two-hundred and four plus one-thousand three-hundred and eighty-one”).

In fact, one of the many subsets of Kanjis is the set of number Kanjis.

Learning the readings of Chinese characters in the Japanese language (from hereon refered to by its romanized name: Kanji) is a complex linguistic challenge, to the point that both psychologists and neuroscientists frequently use Kanji acquisition as the basis for studies focusing on human learning [6] [5].

2.1.1 Kunyomi and Onyomi

Historically, the Japanese language had no writing system until the ancient Chinese civilization introduced it sometime in the 5th century AD [2]. Japanese scholars that mastered the Chinese characters would also learn the Chinese readings, effectively importing Chinese words into the Japanese language. However, native Japanese words for most of the imported vocabulary already existed, meaning acquisition of Chinese loanwords was at first done by the scholars and the elite (not entirely unlike Latin's position in Renaissance Europe [4]). As the Chinese characters became more and more widespread, the concept of ”Kunyomi” and ”Onyomi” developed. Kunyomi means the Japanese reading of a Chinese character, based on the native word for the concept the character represents, while onyomi is the imported pronounciation based on the Chinese language (it must be remarked, however, that onyomi is often wildly different from actual Chinese pronounciation of a character). Since the introduction of kanjis, Japan has gone through several phases of fragmentation and interacting with various parts of East Asia. Sailors from the southern island of Kyushu would interact with people from Canton and the Ryukyu Kingdom (modern day Okinawa), while traders from the north of Japan would interact with people from Manchuria (northern parts of modern day China) and Koryo (modern day Korea). Due to this variation, the Japanese language acquired a wide variety of onyomis for the kanji, and has a much greater variation in pronounciation than either the Korean language or any Chinese dialect. In addition, native Japanese words are subject to dialect differences to the degree that certain words are completely incomprehensible to speakers of respectively a northerner and a south-westerner [3].

2.1.2 Name Readings

Due to these differences, many names that represent the same ideas would use the same kanjis for writing the names, but different ways of pronounciating the names.

3 Article Identification

With Wikipedia offering its articles as a downloadable package, one of the key challenges is to be able to identify relevant articles to use for data sampling.

The first step to solving, is of course to get the data. During the development of the software, only approximately 1/8th of the Japanese Wikipedia was downloaded and extracted. It was fetched from the following URL:

```
http://dumps.wikimedia.org/jawiki/20150602/
jawiki-20150602-pages-articles1.xml.bz2
```

(File size is about 288.8 MB compressed, close to 1.2 GB uncompressed in XML format.)

After the XML filed was exported, WikiExtractor (https://github.com/bwbaugh/wikipedia-extractor) was used to extract the data into more easily handled files. The program was run with the following command:

```
WikiExtractor.py -b 1M -o target jawiki-20150602-pages-articles1.xml
```

The result will then be a set of 1MB files neatly sorted into folders of 100 files each in the directory /target/. The process might take more than five minutes to complete. Once the extraction is complete, the articles are ready to be used as input.

3.1 Tokenizing - Kuromoji

Compared to English, Japanese language is relatively difficult to tokenize due to the lack of whitespaces to separate linguistic entities (words). For reference, the same difficulty is found in natural language processing of other germanic languages such as Norwegian and German where compound nouns are written without any whitespace between the compound words. Japanese is on the rather extreme end, where whitespaces are normally only inserted after punctuation symbols.

For tokenizing Japanese sentences I used Kuromoji, developed by Atilika, which not only is very precise in the separation and tokenization of Japanese words, but also includes syntactic information about the words, such as classifying whether the word is a noun or a verb etc.

A typical Kuromoji output, in this case for the word 物理学 (physics), would look like this:

```
物理	名詞,一般,*,*,*,*,物理,ブツリ,ブツリ
学	名詞,接尾,一般,*,*,*,学,ガク,ガク
```

As demonstrated, Kuromoji splits the three character word into two tokens, with the second one meaning ”the study of”. Following each of the two tokens is a short description. 名詞 means noun, 一般 means regular (commonly used), 接尾 means suffix. The final symbols are guides for regular pronounciations (however, these fields are hardly accurate when dealing with name kanjis).

Kuromoji is included into the project by using Maven. In the project's pom.xml file, the following lines ensured inclusion of the up-to-date version of Kuromoji:

```
<dependency>
    <groupId>com.atilika.kuromoji</groupId>
    <artifactId>kuromoji-ipadic</artifactId>
    <version>0.9.0</version>
  </dependency>
</dependencies>
```

After this, Kuromoji is able to handle input from the Wikipedia article files (after removing the XML-markup from the files).

3.2 Simple Article Analysis

A simple set of analyzing methods that do not require much processing power for filtering out disqualified articles.

3.2.1 Title Analysis

Four methods are used to analyse the title of an article:

1. Title Length Analysis - The simplest of analysis simply checks the length of the title. If the title is a single character, the article will not be about a person. Similarly, titles with more than 8 characters in length will either be about non-persons or persons that do not use Kanjis in their names.

2. Character Set Analysis - Since we are only interested in individuals who use Kanjis in their names, a simple analysis is to check whether the title contains a single kanji. If it does not, the article is disqualified.

3. Keyword Analysis - A set of words that would never be used in a given name. Words include 大学 ("university") and 電車 ("train"). Note: The entire set of Keywords is defined in the file "src/titlekeywords.txt".

4. Title Suffix Analysis - A large amount of kanjis might occur in a persons name, but when used as a suffix will almost never indicate a person. One of the most regular examples is 省, which is often used as the first character of a name, but as a suffix always means "Ministry" (of Defence etc.). Note: The entire set of disqualifying suffixes is found in the file "src/titlesuffixkeywords.txt".

3.3 Token Analyzer

3.3.1 Bag of Tokens

The bag of tokens methods counts the number of occurences of each token in a given text. TODO: Implementere resten, en læremetode som lærer av forekomsten av tokens i artikler som allerede er definert som enten relevante eller ikke. Husk: På dette tidspunktet kan du fint anta at alle artikler som er med videre vil ha kanji i navnet og derfor er det ikke relevant å undersøkte dette, kun innholdet i selve teksten. Husk å filtrere ut unødvendige ord (です、ある、が、を、はetc).

4 Data Extraction

5 Searchable Database

References

[1] Silviu Cucerzan. “Large-Scale Named Entity Disambiguation Based on Wikipedia Data”. In: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (2007).

[2] Bjarke Frellesvig. *A History of the Japanese Language*. Cambridge, 2010.

[3] Nanette Gottlieb. *Language and Society in Japan*. Cambridge, 2005.

[4] Walter J. Ong. “Latin Language Study as a Renaissance Puberty Rite”. In: *Studies in Philology* 56 (1959), pp. 103–124.

[5] Yasuhisa Sakurai et al. “Different cortical activity in reading of Kanji words, Kana words and Kana nonwords”. In: *Cognitive Brain Research* 9 (2000), pp. 111–115.

[6] M. Yamazaki et al. “Two age of acquisition effects in the reading of Japanese Kanji”. In: *British Journal of Psychology* 88 (1997), pp. 407–421.