Charles University

Faculty of Arts

Institute of East Asian Studies

Bachelor Thesis

Ing. Tereza Štefková

Text data mining as an viable method of Japanese studies

Data mining jako metoda použitelná v oblasti japonských studií

Praha 2019

Supervisor: Mgr. Petra Kanasugi, PhD.

Prohlášení

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 7. 5. 2019 Tereza Štefková

Acknowledgment

I would like to express my honest gratitude to my supervisor Dr. Petra Kanasugi for suggesting this interesting topic and her useful comments. I am also very grateful to Kawashima Makiko and Ing. Patrik Urban for the reference summaries.

Last but not least, let me thank my family and friends for their immense support during my studies.

Abstrakt: Tato práce se zaměřuje na problematiku potenciálního využití metod dolování z textu v oblasti japonských studií. První část práce shrnuje základní přístupy dolování z textu a jejich aplikace v praxi. Dále podáváme podrobný výklad problematiky předzpracování textu, u kterého se soustředíme na techniky používané v případě japonštiny a angličtiny.

Hlavní část práce spočívá v aplikaci metod dolování z textu na tři konkrétní výzkumné otázky z oblasti japonských studií. V prvním tématu ukážeme na příkladu děl dvou vybraných japonských proletářských autorů, jak mohou techniky shlukování odhalit zajímavé tematické rysy literárních děl. V případě druhého výzkumného tématu využijeme analýzu sentimentu za účelem vyšetření míry negativního sentimentu, který se objevuje v japonských a zahraničních novinových článcích pojednávajících o návštěvách, které vykonávají japonští představitelé ve svatyni Jasukuni. Nakonec se zaměříme na metody automatického shrnutí dokumentů, které aplikujeme na japonské a anglické texty.

Získané výsledky detailně diskutujeme, zvláště se zaměřujeme na vyhodnocení použitelnosti představovaných metod pro japonská studia.

Klíčová slova: dolování z textu, předzpracování, japonská proletářská literatura, svatyně Jasukuni, automatické shrnutí textu, shlukování, analýza sentimentu

Abstract: In this thesis we address the problem of possible utilization of text mining methods in the field of Japanese studies. We review the fundamental text mining approaches and their practical applications in the first part. Then we elaborate on the topic of preprocessing with special focus on techniques used for Japanese and English texts.

In the main part of the thesis we apply text mining methods to three concrete research questions relevant in Japanese studies. The first research topic illustrates the technique of clustering applied to works written by two Japanese proletarian authors to reveal interesting topic patterns in their writings. The second topic makes use of the sentiment analysis with the aim of studying the extent of negative sentiment expressed in both foreign and Japanese newspaper articles that refer to Japanese officials' visits to Yasukuni shrine. Finally, we address methods of automatic summarization and their application to Japanese as well as English sample texts.

The results obtained are discussed in detail with a special focus on the assessment of viability of the presented methods in Japanese studies.

Key words: text mining, preprocessing, Japanese proletarian literature, Yasukuni shrine, automatic summarization, clustering, sentiment analysis

Contents

Introduction

The amount of text documents available in digital format has been increasing rapidly in the last decades. This enormous volume of a wide range of text data that are being created on social networks or published on websites every day made it impossible for humans to process the data manually. This problem necessitated the development of effective and efficient automatic methods for extraction of meaningful information.

The task of automatic text processing with the aim of extracting useful information is called text mining. This thesis presents an introduction to text mining methods and analyses their potential utilization in Japanese studies.

Even thought we can perceive a trend in the humanities towards their taking advantage of computing and digital technologies, adopting text mining methods is by no means a common approach in the field of Japanese studies. The main purpose of this thesis is to apply several text mining techniques to three different research problems and assess whether text mining yields relevant results or not. The research problems address questions related to Japanese literature, society, history.

This thesis is organized as follows: In the first chapter we start with a brief definition of text mining and related terms. We also review fundamental methods that represent the most fruitful parts in the field of text mining. Having introduced the methods, the focus is shifted to the practical applications.

The second chapter is devoted to the problem of preprocessing, which represents a common starting point for most of the text mining methods. We discuss the steps in the preprocessing in detail and illustrate the techniques with help of example sentences. Since we analyze both Japanese texts and English texts in the thesis, we comment on the significant differences in the Japanese and English preprocessing techniques.

Having provided an introduction to text mining methods and preprocessing, we move to three specific research topics. In Chapter 3 we address the problem of Japanese proletarian writings by Kuroshima Denji and Miyamoto Yuriko. We use the techniques developed for clustering with the aim of identifying groups of works with similar topics for the respective cases of the two authors.

In the fourth chapter we aim at answering the research question, whether foreign newspaper agencies tend to express more negative sentiment in articles related to Japanese officials' visits to the controversial Yasukuni shrine. To this end we employ methods provided by sentiment analysis.

The last research problem deals with the automatic summarization techniques. We demonstrate these techniques on two English-written articles and two Japanese-written articles and evaluate the performance.

We conclude with a detailed discussion of viability of the methods presented in this thesis for research in the field of Japanese studies and suggest modifications that could be adopted to enhance the performance of these method.

Note that the Japanese names are listed in the format where the surname precedes the given name, as is common in Japan. The English translations are presented for all Japanese terms at the first occurrence in the thesis.

Chapter 1

Introduction to text mining

The aim of this chapter is to provide an overview of basic concepts of text mining. While the first part focuses on a wide range of text mining methods, the second part deals with its applications to concrete problems.

Text mining is defined as the analysis of text data with the aim of extracting meaningful information useful to the person performing the analysis [1]. According to [2], one characteristic feature of text data is their lack of structure (they represent the so-called unstructured or semi-structured data). This is in stark contrast with the structured data which are comprised of predefined data types organized and stored in a highly structured format of a database. The process of finding useful patterns in these data is called data mining – an analogue to text mining.

Before proceeding further, let us clarify several core terms that will appear thorough this thesis. Adopting definitions stated in [3], a document can be any relevant segment of text (a sentence, a paragraph, a chapter, an e-mail...). In this thesis we use terms document and text interchangeably, since in our analysis we deal either with the whole literary work or an article. A collection of documents is called a corpus while collection of all unique words appearing in the corpus are referred to as lexicon. A term is usually a word or a phrase, whereas by token we understand any meaningful segment of a sentence such as words, phrases or symbols [4]. We use the expressions term, word and token interchangeably in those parts of the thesis where it would not be misleading.

1.1 Methods

When performing text mining on a document or a corpus, we employ one or more text mining methods. In this section we review the main approaches utilized in the field of text analysis.

Each of the methods makes use of a wide range of algorithms that can vary according to the type of texts we analyze or the precision we want to achieve. The algorithms can be classified as supervised and unsupervised [1]. Unlike unsupervised algorithms, supervised algorithms require certain amount of learning data that were processed manually to learn the task needed. The algorithm can be afterwards applied to evaluate the test data.

1.1.1 Information retrieval

The task of information retrieval is to find text documents from a corpus that contain information of interest. Thus, information retrieval is more concerned with data access and its collection than with the analysis itself. Nowadays, the information retrieval algorithms are being widely used by search engines on the World Wide Web [5].

1.1.2 Information extraction

Information extraction deals with automatic extraction of structured information from textual data. The structures to be extracted are typically so-called named entities (sequences of words that refer to some real entities) and the relations between them. The corresponding core tasks of information extraction are then referred to as entity recognition and relation extraction. The former focuses on identification of named entities in the text and their matching with the predefined set of named entity types. The most common types are person, organization and location, nevertheless different types can be defined for specific domains. Once the named entities are detected and classified, the relation extraction algorithms, which identify and characterize the semantic relation between them, can be employed.

Let us illustrate the process and terms defined above on a simple example:

Mark Zuckerberg founded Facebook.

We can detect two named entities and label them, namely

Mark Zuckerberg[person], Facebook[organization].

The relationship between them can be denoted as

FounderOf[Mark Zuckerberg, Facebook].

Information extraction methods use both supervised and unsupervised algorithms to extract and label named entities and the relations. Information extraction is usually

utilized in combination with other text mining methods such as text summarization, clustering, classification etc., since they facilitate further processing of the text by adding structured information to it [1].

1.1.3 Natural language processing

Natural language processing is a part of computer science which employs computers to analyze, represent and manipulate natural text [6]. The natural language processing techniques can be divided into the following categories: natural language understanding, natural language generation, speech or voice recognition, machine translation, spelling correction and grammar checking. A wide range of natural language techniques are useful in many text mining problems such as part-of-speech tagging or tokenization.

1.1.4 Text summarization

The goal of text summarization is to provide a good-quality summary of a given text or a set of texts, as stated in [7], [1]. We distinguish between the so-called extractive summarization and abstractive summarization. The former is unification of certain parts (mainly sentences) extracted from the input text that convey relevant information. The latter approach which reproduces the core information in a new way, requires advanced language generation techniques. Therefore, most of the algorithms are based on extractive summarization. We will further elaborate on the topic of text summarization in Chapter 5.

1.1.5 Clustering, topic modeling

Clustering divides objects into groups according to some similar features. The similarity is determined by prescribed similarity function. The problem of clustering can be utilized in many areas such as document organization, browsing, generation of corpus summarization or text classification. Topic modeling, on the other hand, is a softer version of clustering, in the sense that we only specify probability for each text to belong to certain topic. This topic is discussed in detail in [1] and [8].

Both clustering and topic modeling are examples of so-called unsupervised learning, since they do not require training. In Chapter 3 we provide more detailed overview of clustering.

1.1.6 Classification

Unlike clustering, classification encompasses various supervised method. The learning set contains documents manually labelled according to the class they belong to. Based on the comparison with the learning set data, the algorithm then assigns the test document to one of the classes. The classification algorithms can be divided into two parts: the hard version classification, in which case the test data is assigned to one class and the soft version classification, where we obtain the probabilities with which the document belongs to the respective classes [1].

1.1.7 Sentiment analysis

The goal of sentiment analysis is to detect sentiment and opinion expressed in the text and classify it according to the polarity of the sentiment. It deals with two basic problems: detecting subjectivity in texts and evaluating the polarity of the subjective part of the text, i.e. deciding whether the sentiment is positive or negative. We distinguish between two fundamental sentiment analysis approaches: machine learning approach, including both supervised and unsupervised methods, and lexicon-based approach [9]. In Chapter 4 we describe the latter approach in detail.

1.2 Applications

Text mining is not a mere theoretical concept, it has a wide range of practical applications. In this section we briefly summarize the most prominent fields that make use of text mining methods.

One of the most prominent users of text mining methods is the biomedical science. Recently, the number of biomedical publications have risen significantly, making it necessary to process the texts automatically. This field makes use of for example the information extraction to detect and classify named entities often represented by proteins, diseases or genes, to name a few. This is particularly difficult as there exist a lot of synonyms and abbreviations for the given term. The relation extraction, on the other hand, aims at finding relations between the named entities, for example the relationship between a gene and a disease. The information extraction is usually used as a starting point for summarization of the biomedical publications [7].

Another established field for text mining is the social media. People from every walk of life produce a huge amount of texts every day using social media such as Facebook or Twitter. Social media posts and comments can provide material for sentiment analysis

of people's reactions to certain events. The characteristic feature of these texts is their length and often poor language that requires a lot of normalization.

Academic publications also represent a natural source for text mining. This field successfully takes advantage of clustering methods to organize articles according to the topic they address or text summarization methods to identify the key parts of articles. This can be achieved with use of the so-called impact summarization, in which case a corpus of articles is needed. The algorithm then identifies the articles that cite the analyzed article and extracts the parts concerning the citation. This method is based on a simple assumption that the context in which the original paper is cited contains crucial information about the summarized paper. The resulting summary is comprised of sentences that are similar to sentences appearing in the later papers at the places of reference to the original paper [1].

Text mining techniques are also often used in business intelligence. In this case, text mining helps companies to predict their customers' needs and to map their opinion about the product they offer.

Chapter 2

Preprocessing

This chapter addresses the problem of text preprocessing which plays a vital role in the majority of text mining applications. Since this thesis deals with text mining of both English documents and Japanese documents, we comment on the significant differences in the preprocessing stemming from the distinctive characteristics of both languages.

The preprocessing is more challenging in the case of Japanese language owing to several reasons. Firstly, words are not clearly separated by spaces as is the case of English, making the detection of word boundaries a demanding task. Secondly, Japanese makes use of four possible scriptive representations of words, namely the Chinese characters kanji, syllabaries hiragana and katakana and Latin scripts romaji. As a result, the same word can acquire several forms thorough the same document, as discussed in [10].

Preprocessing is usually comprised of the following steps: cleaning of the text, its subsequent tokenization, normalization and filtering of tokens that do not convey factual information. The preprocessed text then acquires one of possible forms on which algorithms can be used. In the following sections we discuss in detail the process of transforming the raw text to some meaningful representations. Detailed explanation of the preprocessinig technique is provided in [4] and [8].

2.1 Preprocessing

In this section we focus on the so-called preprocessing which is a transformation of an unstructured text to a certain kind of representation. The preprocessed text can be further analyzed using various text-mining methods and algorithms.

We will illustrate the steps of preprocessing on both English and Japanese example sentence:

2 years ago, I decided to move to Norway with my sister Anna.
2年前、私はアンナという妹と一緒にノルウェイに引っ越すことにしました。

For tokenization, we use predefined functions from Natural Language Toolkit (NLTK) library for English written text and MeCab tokenizer in the case of Japanese texts. Both of these packages can be readily used in Python, programming language in which the scripts used for analysis in this thesis are written.

2.1.1 Cleaning

The biggest source of unstructured text data nowadays is the World Wide Web. Since the text mining methods usually aim at processing large amount of data, it is necessary to automatize the process of data collection. However, the unstructured data in the HTML format downloaded from the Internet are usually burdened by markups (characters that encode, how the text should be displayed in a browser). Therefore, the first stage of preprocessing involves text cleaning, which excludes the redundant characters from the HTML format source.

2.1.2 Tokenization and part of speech tagging

Having obtained a text in a form comprehensive for humans, we can proceed to tokenization. In this process the text, which is a string of characters, is converted into a list of tokens. Unlike in Japanese, in most languages the words are separated by spaces in most of the languages which facilitates the process of tokenization.

We use the NLTK library tokenizer and part of speech tagger to identify the tokens in the English example sentence. The result then attains the form of a list of tuples consisting of the token and the corresponding part of speech:

[('2', 'CD'), ('years', 'NNS'), ('ago', 'RB'), (',', ','), ('I', 'PRP'), ('decided', 'VBD'), ('to', 'TO'), ('move', 'VB'), ('to', 'TO'), ('Norway', 'NNP'), ('with', 'IN'), ('my', 'PRP$'), ('sister', 'NN'), ('Anna', 'NNP'), ('.', '.')]

NLTK tokenizer recognizes 36 different part of speech tags. In the above example, the part of speech tags' abbreviations stand for: CD cardinal digit, NNS noun plural, RP particle, PRP personal pronoun, VBD verb past tense, TO to go 'to', NNP proper noun, singular, IN preposition/subordinating conjunction, PRP$ possessive pronoun,

NN noun singular.

Even though the tokenization of Japanese texts is more challenging, several methods have been developed so far. The major approach is based on a dictionary of tokens. The program then tries to identify the predefined tokens in the text we analyze and segment the document accordingly.

One of the most common programs for Japanese tokenization based on dictionaries is MeCab that was developed at Nara Institute of Science and Technology (奈良先端科学技術大学院大学). Apart from tokenization, this text segmentation program also provides the part of speech information about the tokens. See [11] and [12] in order to gain deeper insight into Japanese morphological analysis and the techniques used by MeCab.

Let us illustrate the performance of MeCab on the example sentence. When tokenizing text, MeCab can be used in two modes: Wakati which simply adds spaces between tokens and Chasen that also provides the information about the reading of the kanji and the corresponding part of speech category. The result for both Wakati and Chasen mode are displayed below and in Table 2.1.2, respectively.

２ 年 前 、 私 は アンナ という 妹 と 一緒 に ノルウェイ に 引っ越す こと に し まし た 。

2.1.3 Normalization

Obviously, not every token conveys meaningful information and some of the tokens differ only in the way they are incorporated in the sentence (for example upper case letters at the beginning of the sentence) or in their representation (depending on the author, words can be written in both kanji and hiragana).

To reduce the number of tokens, we can perform certain transformations of the tokens described in [3] and [8]. One of them is the so-called stemming during which the suffixes and prefixes are removed and only the stem of a word is taken into consideration. For instance, stemming process reduces the words ”suggests” and ”suggested” to the dictionary form ”suggest”. However, the stemming does not always lead to accurate results and lacks precision. This becomes apparent on the example of words ”relativity” and ”relation”, which would both result in a single token ”relate” despite their different meaning.

A method alternative to stemming is the so-called lemmatization that aims at mapping the tokens to their dictionary forms. Unlike in case of stemming, in lemmati-

Token	Reading	Dictionary form	Part of speech
2	ニ	2	名詞-数
年	ネン	年	名詞-接尾-助数詞
前	マエ	前	名詞-副詞可能
、	、	、	記号-読点
私	ワタシ	私	名詞-代名詞-一般
は	ハ	は	助詞-係助詞
アンナ	アンナ	アンナ	名詞-固有名詞-人名-一般
という	トイウ	という	助詞-格助詞-連語
妹	イモウト	妹	名詞-一般
と	ト	と	助詞-格助詞-一般
一緒	イッショ	一緒	名詞-サ変接続
に	ニ	に	助詞-格助詞-一般
ノルウェイ	ノルウェイ	ノルウェイ	名詞-一般
に	ニ	に	助詞-格助詞-一般
引っ越す	ヒッコス	引っ越す	動詞-自立五段・サ行基本形
こと	コト	こと	名詞-非自立-一般
に	ニ	に	助詞-格助詞-一般
し	シ	する	動詞-自立サ変・スル連用形
まし	マシ	まし	助動詞特殊・マス連用形
た	タ	た	助動詞特殊・タ基本形
。	。	。	記号-句点

Table 2.1: Tokenization results of the example Japanese sentence using Chasen mode.

zation the reduction is performed systematically taking into consideration the part of speech of each word which results in higher precision in distinguishing tokens conveying different meaning (tokens ”relativity” and ”relation” would be treated as distinct words).

Note that besides the application of lemmatization and stemming is also desirable to lower the letters and unify the representation of numbers.

For our sample sentences the lemmatization yields the following list of tokens:

two year ago i decide to move to norway with my sister anna

２ 年 前 私 は アンナ という 妹 と 一緒 に ノルウェイ に 引っ越す こと に する ます た

Notice that in the case of the Japanese sentence, the lemmatized tokens correspond to the dictionary form outputs from Chasen. The Chasen mode in MeCab enables us to perform lemmatization readily for Japanese texts.

2.1.4 Filtering

Next step in the preprocessing is the removal of tokens that do not contribute to the information the text conveys. This involves exclusion of redundant tokens such as punctuation and the so-called stop-words. On one hand, these are words that convey little information such as prepositions, articles or auxiliary verbs even though they appear very frequently in the text. It is estimated that the number of stop-words can amount to $20-30\%$ of the overall number of tokens [4]. On the other hand, it is usually assumed that rare words are also insignificant for our understanding of the text.

There are two ways how to determine the list of stop-words – we can either use a predefined list, or generate our own list based on the lexicon in the corpus we analyze. The second method is more suitable when dealing with specific types of texts. The stop-words then contain the words with the largest number of occurrences and the rare words.

The stop-words removal represents an essential part of text preprocessing, since it significantly reduces the number of tokens while excluding redundant words which present obstacles in further analysis. This is due to the fact that we usually attribute more importance to tokens with high occurrence in the text. To obtain meaningful results, we need to eliminate stop-words so that they would not interfere with the words relevant to the topic of the text.

The stop-words removal in the example sentences yields:

year ago decide move norway sister anna

年 前 アンナ 妹 一緒 に ノルウェイ 引っ越す

2.2 Text representation

Having performed the preprocessing of the given text, we need to take one last step towards the creation of its meaningful representation which facilitates the process of further analysis.

Ideally, we would like to obtain a representation that would respect the semantics of the original text. This can be partly achieved by taking into consideration named entities and the relations between them. However, it is very demanding to obtain a good quality semantic representation of a text. As a result, one of commonly used representations is the so-called bag of words representation [8]. This model ignores the relations between the respective tokens as well as their order and considers only the number of their occurrences.

According to [8] the most commonly used representation based on the bag of words scheme is the Vector space model that treats the preprocessed text as a vector (a column of numbers) which dimension (i.e. the number of elements) is the overall number of distinct tokens (known as types in linguistics). Each of the elements (an entry of the vector) then carries information about the number of occurrences of the given token in the text and is usually called the weight. In other words, we can think of this vector as a list of tokens with the corresponding weights. The weight itself can be defined in several ways according to the concrete problem we study. The simplest example is that of a binary representation, in which case the weights attain values 0 or 1 according to whether the term appears in the document or not. However, in most of the applications more refined weighting schemes are required such as the TI-IDF method applied when we analyze corpora (for details see Section 3.2.1).

In real world applications we usually work with a set of documents – a corpus. The vector space representation of a corpus is a natural extension of the vector representation concept – a corpus is represented by a matrix (a rectangular array of numbers) called the term - document matrix. Let us assume that we have d distinct tokens in the corpus that consists of n documents. Under these assumptions, the term-document matrix has d rows corresponding to the terms and n columns that refer to the respec-

tive vectors of the documents. The dimension (i.e. the number of elements) of the resulting term-document matrix is the product of n and d. Its elements then carry the information about the number of occurrences of the given type in the concrete document – the element in the d-th row and n-th column corresponds to the weight of the d-th term in the n-th document.

The text data represented by the Vector space model have several characteristic features, as described in [1]. Firstly, they are high dimensional, in the sense that the dimension of the term-document matrix can be extremely large for big corpora (stemming from large number of documents) with an extensive lexicon (leading to a large number of types). The second typical property of the Vector space model representation is its sparsity, meaning that a large number of term-document matrix elements are equal to zero. The sparsity can be attributed to the fact that not every document contains all tokens appearing in the corpus. Moreover, the length of the documents may vary considerably throughout the corpus which also contributes to the sparsity of the Vector space model and necessitates normalization.

This rather simple Vector space representation of the texts enables us to employ a wide range of mathematical methods. Several techniques will be introduced throughout this thesis.

Chapter 3

Text mining of Japanese proletarian literature

This chapter is devoted to potential usability of clustering methods to Japanese studies. We apply of the clustering techniques – Latent semantic analysis – to Japanese proletarian writings and discuss the results in detail.

3.1 Research problem and methods

Clustering methods usually require a large corpora; unfortunately only a minority of Japanese corpora are accessible online and free of charge. One of the freely available corpora is the Japanese digital library Aozora bunko (青空文庫, accessible on the webpage https://www.aozora.gr.jp/) which is a collection of more then 15000 literary works from authors whose copyrights have expired.

In this chapter we apply clustering techniques to literary works written by Japanese proletarian writers with the aim of detecting interesting patterns and revealing the topics depicted in their works. The analysis was restricted to two proletarian authors, namely Kuroshima Denji and Miyamoto Yuriko, whose works are freely accessible from the database of Aozora bunko.

We resolved to use of so-called Latent semantic analysis, which is one of a wide range of techniques which clustering offers. To evaluate the quality of the results obtained, we make a comparison with literary studies of the above mentioned authors.

The structure of the chapter is as follows: in the first part we review the methods of clustering with the emphasis on the Latent semantic analysis. Then we introduce the two proletarian writers and their works. Finally, we present the concrete utilization of the Latent semantic analysis in our research problem and discuss the results.

3.2 Clustering – methods

Clustering belongs to extensively studied and immensely useful methods of text mining. Generally, clustering is defined as a process of finding groups of similar items in the given data. In the context of text mining we aim at organizing the documents (or paragraphs, sentences etc.) in our corpus into sets sharing some common feature.

In Chapter 2 we already commented on the fact that text data are sparse and high-dimensional. In the clustering process we usually aim at the dimensionality reduction. In the following we introduce the so-called Latent semantic analysis (LSA) which is a widely used method for text clustering based on the dimensionality reduction.

3.2.1 TI-IDF weighting

The LSA algorithm requires some of the vector space representation which differ in the weighting scheme – a method of assigning importance to the respective tokens. One of the commonly used method is so-called Term frequency-inverse document frequency (TI-IDF) which transforms the vector representing one text to a vector of the same dimension with different weights. The formula for the weights reads

$$w(t) = f_d(t)\log\frac{D}{f_D(t)} \tag{3.1}$$

where $f_d(t)$ is the frequency of the token t in the text d, D is the number of documents and $f_D(t)$ is the frequency of the token t in the corpus. The TI-IDF weighting relies on the assumptions that words appearing often in the corpus are not very significant for the information the text conveys (TI-IDF can detect stop-words). The algorithm tends to assigns larger weights to terms that occur rarely in the corpus while being very frequent in the given document. In other words, it assigns large weights to words that have potential to bear information about the topic of the document [1].

3.2.2 Latent semantic analysis

One of the algorithms that reduce the dimensionality of the vector space representation of the corpus is so-called Latent semantic analysis (LSA) [1] that is based on the principle that terms used in similar context have similar meanings. LSA offers a method to alleviate the problems of synonymy and polysemy that the Vector space model is unable to cope with. This algorithms produces a set of concepts (corresponding to topics in our analysis) that appear in the documents and evaluates the relationships between the texts and these concepts.

Latent semantic analysis is based on so-called singular value decomposition, according to which the term-document matrix representing the corpus can be decomposed into a product of three matrices. This decomposition enable us to reduce the dimension of the term-document matrix by replacing certain elements by 0. Thus, LSA identifies the most important concepts in the corpus and similarity of the documents with these concepts. More technical explanation of the principles of singular value decomposition and LSA are far beyond the scope of this thesis; for more details we refer to [5].

3.3 Japanese proletarian literature

The proletarian literature appeared in Japan at the beginning of the 1920's. Spurred by the spreading of the ideology of Marxism, a large number of authors reflected in their writing the bad conditions of the working class people. Japanese literary scene abounded with proletarian literature in the years between 1925 and 1935, but started to decline in 1935 when many of the proletarian writers were forced to renounce their ideology which was banned by the Peace preservation law. The literature produced by those who seceded from the proletarian literature is referred to as the tenkō bungaku, literature of conversion [13].

In the analysis we focus on two outstanding proletarian authors, namely Kuroshima Denji and Miyamoto Yuriko.

3.3.1 Kuroshima Denji

Kuroshima Denji (黒島伝治, 1898-1943) undoubtedly belongs to the best Japanese proletarian writers. He was born into a family of farmers and worked in a factory after completing the vocational school. In 1921 he was sent to Siberia during the so-called Siberian Intervention, but due to his constantly worsening tuberculosis, he was allowed to return to Japan where he joined the proletarian literature movement [13].

Between years 1925 and 1932 he published approximately 60 stories. His writings were spurred by two main sources of inspiration [13]:

1. His experience from the military intervention in Siberia: 雪のシベリア (Siberia in the snow, 1928), 渦巻ける烏の群 (A flock of swirling crows, 1928), 橇 (The sleigh, 1927)

2. Rural life in his hometown: 豚群 (A herd of pigs, 1926), 二銭銅貨 (The two-sen copper coin, 1926), 電報 (The telegram, 1925), 砂糖泥棒 (The sugar thief, 1923), 浮動する地価 (Land rising and falling, 1926),

3. Other works: 穴 (The hole, 1928), 武装せる市街 (Militarized streets, 1930)

His novel 武装せる市街 stands out among his writings not only owing to its length (it is his longest story), but also due to the controversial topic it depicts. It is a story conveying the events preceding the Second Sino-Japanese war which based on his actual research in China.

3.3.2 Miyamoto Yuriko

Miyamoto Yuriko (宮本百合子, 1899-1951), one of the most prominent proletarian writers, began writing her fist novel 農村 (A farm village, 1915) under the influence of the Shirakaba school at the age of 16. She spend several years both in the United States and the Soviet Union; the travels proved to be a rich source of inspiration for her writings. She represented a leading figure in the proletarian literature movement which led to several imprisonments for her political views, a topic that is repeatedly reflected throughout her works [13]

She married twice, her relationships and their failures are portrayed in many of her autobiographical or semi-autobiographical novels such as the famous works 伸子 (Nobuko, 1924) or 播州平野 (The Banshū plain, 1946). Unlike many other preletarian writers, Miyamoto Yuriko belongs to those who refused to bow to the Japanese political military regime and continued writing proletarian works during the war as well as in the post-war period.

Some of her most accomplished writings are listed below. According to [14] and [15], these works can be grouped as follows based on the similarity of topics and motifs:

1. Peasant life: 農村 (A farm village, 1915), 貧しき人々の群 (A crowd of poor people, 1916), 禰宜様宮田 (Mr. Miyata, a Shinto priest, 1917), 風に乗って来るコロポックル (Koropokkuru riding the wind, 1918)

2. Semi-autobiographical writings about her relationships

    - unhappiness in her first marriage: 我に叛く (Against me, 1921), 火のついた踵 (The burning heel, 1922), 心の河 (Rivers of the soul, 1924), 伸子 (Nobuko, 1924)
    - her second marriage and separation from her husband during his imprisonment: 播州平野 (The Banshū plain, 1946), 風知草 (Purple grass, 1946)
    - sequels to 伸子: 二つの庭 The two gardens, 1947), 道標 (Milestones, 1947)
    - about her relationship with a woman: 一本の花 (A flower, 1927)

3. Experience from the Soviet Union: 新しきシベリアを横切る (Crossing the new Siberia, 1931), ズラかった信吉 (The runaway Shinkichi, 1931), ペーチャの話 (Pēcha's story, 1931), おもかげ (Images, 1940), 広場 (The square, 1940)

4. Her imprisonment: 一九三二年の春 (The spring of 1932, 1932), 刻々 (Moment by moment, 1933)

5. Political fiction criticizing the proletarian movement: 小祝の一家 (The family of Koiwa, 1934), 乳房 (The breast, 1935)

3.4 Clustering of proletarian literature – results

Let us now perform the clustering on the selected works by Kuroshima Denji and Miyamoto Yuriko. We applied LSA method to 10 writings by Kuroshima Denji (for reference see Table 3.1) and 22 works written by Miyamoto Yuriko (Table 3.2).

The analysis was carried out in Python, the LSA algorithm was imported from the predefined Gensim module (Gensim homepage https://radimrehurek.com/gensim/). The steps of the algorithm are as follows:

1. Read the documents in the corpus

2. Perform the preprocessing on the corpus (to reduce the lexicon used in the corpora we opted to consider only nouns in the analysis and discard other tokens)

3. Assign the TI-IDF weights

4. Perform LSA

The LSA algorithm in Gensim has one parameter that needs to be specified – the number of topics. Once this number is set, the algorithm yields the specified number of topics with the corresponding topic words ordered from the most significant to the least significant. In addition to this, we also obtain weights for the documents that characterize how much the given document corresponds to the given topic.

However, the interpretation of the concrete results can be slightly challenging, since both the topic words and the document weights can attain negative values. We will shed some light on this problem in the analysis of the selected proletarian writings.

First, we process the selected works written by Kuroshima Denji. The number of clusters was set to 2 since we aim at detecting two major topics reflecting in Kuroshima's stories – the Siberian experience with Japanese military and rural life in remote parts of Japan.

The LSA method yielded the following two topics characterized by the topic words (ordered from the most important one). Notice that these attained both positive and negative score, as becomes apparent in the case of the second topic. The topic words can be displayed together with the numerical value of the score which corresponds to the significance of the given topic word with respect to the topic. However, we resolved not to write the score explicitly, for the sake of clarity.

Topic words 1:
positive score: 雪 (snow), 中隊 (troop), 銃 (gun), 橇 (sledge), 丘 (hill), 兎 (rabbit), シベリア (Siberia), 兵士 (soldier), 兵卒 (private soldier), 戦争 (war), 支那 (China), 少佐 (lieutenant), 大隊 (battalion), 商人 (merchant), パン (bread), 老人 (old person), 内地 (inland), 病院 (hospital), 防寒 (protection against cold), 曠野 (wasteland)
negative score: none

Topic words 2:
positive score: 杜氏 (chief brewer at a sake brewery), 豚 (pig), 醤油 (soy sauce), 親爺 (boss), 主人 (master), 柵 (fence), 支那, 中学 (middle high school), 坊っちゃん (son), 田 (rice field), 敷地 (site), 田畑 (fields), 鶴 (crane), 土地 (soil), 砂糖 (sugar), 小作 (tenant farming), おふくろ (bag), 息子 (son), 小屋 (hut)
negative score: 雪

Title	Topic 1	Topic 2
砂糖泥棒	0.11	0.50
電報	0.11	0.43
豚群	0.22	0.63
二銭銅貨	0.03	0.09
浮動する地価	0.19	0.50
橇	0.74	-0.22
雪のシベリア	0.67	-0.16
渦巻ける烏の群	0.79	-0.20
穴	0.25	0.04
武装せる市街	0.28	0.26

Table 3.1: The selected works by Kuroshima Denji listed according to the year in which they were written. The right part of the table displays the weights of the writing corresponding to the 2 topics defined by the set of topic words listed above.

From the above topic words follows that the first topic is more related to the Siberian intervention whereas the second one contains central terms related to the rural oriented stories. This result is in accordance with our expectations, since literary studies identified those two topics in Kuroshima's writings.

Together with the above topic words we obtained the LSA weights for the documents (see the columns "Topic 1" and 'Topic 2" in Table 3.1). Two fundamental questions now arise: how should we interpret the positive/negative scoring of the topic words and what is the relation between this scoring and the weights assigned to the documents?

The weights (corresponding to one specific topic for each document) reflect the number of occurrence of the topic words in the text, their significance for the topic and their polarity. In other words, the weight of the given document corresponding to a specific topic is a sum over the topic words appearing in the text normalized so that the resulting weight is in the range from -1 to 1. The topic words for a document with a weight close to 1 are very likely to capture the topic described in the document well. On the other hand, negative value of the weight suggests that the text is characterized by the topic words with negative scores.

Having clarified the topic words scoring and the interpretation of the weights, let us now turn to the interpretation of the results shown in Table 3.1. The Siberia-related works (for reference see Section 3.3.1) achieved higher score in weighting (see Table 3.1) since they contain a large number of Topic words 1. On the other hand, texts related to the rural village life scored higher in the Topic 2. Moreover, The Siberia-oriented works attained negative weights with respect to Topic 2 since they have a huge occurrence of the negative topic word " 雪". For each of the two topics we marked by red color the works that are most similar to the given topic (Table 3.1).

The analysis strongly supports the hypothesis that the selected works written by Kuroshima can generally be classified into two groups: those depicting his military experience in Siberia and the rural ones. However, the analysis is inconclusive in case of two writing: 武装せる市街 and 二銭銅貨 that do not show any decisive relation to Topic 1 or Topic 2.

Now, we will shift the to the works by Miyamoto Yuriko. We selected the most prominent writings from every stage of her literary career (for details we refer to Table 3.2). The selected works are also listed and classified according to themes in Section 3.3.2.

The number of topics was set to 7. The topics with the corresponding topic words are listed below:

Topic words 1:
positive scoring: 母 (mother), 良人 (husband), 留置 (imprisonment), 祖母 (grandmother), こころ (soul), 結婚 (marriage), モスクワ (Moscow), プラット (Pratt), ミス

(mistake), 夫人 (husband), 愛 (love), 婦人 (woman), 監房 (cell, ward), 文学 (literature), 看守 (jailer)
negative scoring: none

Topic words 2:
positive scoring: 留置, 監房, 看守, サークル (club), プロレタリア (proletarian), 主任 (person in charge), 同盟 (league), 警察 (police), 文化 (culture), 革命 (revolution), ソヴェト (Soviet)
negative scoring: 良人, 母, 併 (but), 愛

Topic words 3:
positive scoring: 祖母, 海老屋 (Ebiya), 村 (village), ペーチャ (Pēcha) , 善 (goodness), 婆 (grandmother), 豊 (abundant), 餅 (rice cake), 婆さん (grandmother), 乙女 (maiden), 菊 (chrysanthemum), 年寄り (old people)
negative scoring: 良人, 母, 留置

Topic words 4:
positive scoring: 留置, 監房, 看守, 良人, サークル, 主任, 海老屋, 祖母
negative scoring: モスクワ, パリ (Paris), ウラジヴォストク (Vladivostok), 列車 (train), ソヴェト, こころ

Topic words 5:
positive scoring: ペーチャ, 良人, モスクワ, ウラジヴォストク, 農場 (farm), サヴェート (Soviet), 集団 (group), 裁判官 (judge), ソヴェト, グリーゼル (given name), ヌック (given name), ヤーシャ (given name), 併, シベリア (Siberia)
negative scoring: こころ

Topic words 6:
positive scoring: 乙女, モスク, 露台 (balcony), パリ, ヴェルデル (name), 公園 (park), 下宿 (lodging), 留置
negative scoring: 託児 (day nursery), 風知草 (Hakonechloa macra), こころ, 応援 (aid), 新道 (new road), 後家 (widow), リュック (given name)

Topic words 7:
positive scoring: 乙女, 祖父, 祖母, 祖, 託児, ペーチャ, 父ちゃん, 岬 (cape), 掻巻 (type of clothing), 焼酎 (liquor)
negative scoring: 海老屋, 豊, 年寄り, 番頭 (clerk), アイヌ (Ainu)

It can be clearly seen that it is a common feature for the topic words to belong to more than one topic. In such a case the positive/negative topic words from the previous topics are redistributed to form a new topic.

The corresponding weighting of the documents is displayed in Table 3.2. For each of the works we selected the weight which was the larges in the absolute value (ignoring the polarity of the weight). This weight is marked by red color. This process led to the creation of 7 groups. The first group contains the writings that scored highest in Topic 1 (我に叛く, 火のついた踵, 心の河, 伸子, 一本の花, 広場, 播州平野, 風知草, 二つの庭, 道標). The next 3 groups correspond to the most prominent writings related to Topics 2, 3 and 5. Notice that Topic 4 seems to be somewhat redundant. The interpretation is more complicated in the case of Topic 6, where the clustering marked texts with both positive and negative weights. The large positive score for おもかげ can be understood as its high relatedness to positive-score Topic words 6 whereas the negative weight of 乳房 can be attributed to its similarity with the negative topic words corresponding to the sixth topic. Obviously, text 小祝の一家 shows little resemblance to any other writings and stands alone in the group defined by Topic 7.

However, there are other conclusions to be inferred from the Table 3.2. A closer observation of the data reveals further interesting patterns. The writings that belong to the same group also tend to attain similar weights in relation to other topics which is another evidence supporting the hypothesis that texts in the same group possess a strong similarity to each other. This is evident for instance in the case of works 播州平野 and 風知草 that attained similar weights throughout the topics.

The analysis shows results that correspond very well to the classification presented in Section 3.3.2 with some minor deviations represented by works おもかげ and 広場 that did not show similarity with Topic 5, in contrast to our expectations.

The results presented above and their comparison with literary studies show a strong evidence that automatic unsupervised clustering methods could potentially help us to detect various topics depicted in literature and construct groups of writings sharing the same or similar themes.

However, the LSA technique proved not to be the optimal method for soft clustering owing to demanding assessment of the results. For a purpose of future analysis we suggest using the so-called Latent Dirichlet allocation which is a method extensively used in soft clustering [1]. The biggest merit of this method is a straightforward interpretation of the results as Latent Dirichlet analysis yields topic words for predefined number of topics together with the probabilities for the documents to belong to a given topic. This is in stark contrast to LSA, in which case we obtain weights that can attain

both positive and negative values leading to a somewhat complicated interpretation of the results.

Title	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
農村	0,33	-0,13	0,52	0,16	-0,15	0,10	0,10
貧しき人々の群	0,33	-0,16	0,50	0,23	-0,15	0,10	-0,05
禰宜様宮田	0,24	-0,13	0,41	0,18	-0,09	0,02	-0,27
風に乗って来るコロポックル	0,21	-0,16	0,28	0,21	0,07	-0,01	-0,28
我に叛く	0,44	-0,37	-0,28	0,22	0,24	-0,01	0,00
火のついた踵	0,33	-0,32	-0,26	0,27	0,30	0,04	-0,03
心の河	0,41	-0,35	-0,29	0,21	0,23	-0,06	0,08
伸子	0,60	-0,23	-0,08	-0,01	-0,09	0,09	0,08
一本の花	0,38	-0,01	-0,05	0,01	-0,08	-0,12	0,16
新しきシベリアを横切る	0,24	0,20	0,23	-0,32	0,50	-0,09	-0,01
ズラかった信吉	0,28	0,28	0,21	-0,21	0,49	0,10	0,02
ペーチャの話	0,07	0,08	0,23	-0,07	0,42	-0,04	0,15
一九三二年の春	0,32	0,65	-0,15	0,40	-0,03	0,13	-0,06
刻々	0,34	0,65	-0,13	0,43	-0,02	0,12	-0,04
小祝の一家	0,09	0,01	0,14	0,06	-0,12	0,15	0,85
乳房	0,25	0,21	0,03	0,03	0,03	-0,38	0,16
おもかげ	0,29	-0,04	-0,10	-0,31	-0,12	0,49	-0,08
広場	0,34	0,02	-0,05	-0,27	-0,06	0,20	-0,11
播州平野	0,47	0,04	0,10	-0,21	-0,12	-0,45	-0,06
風知草	0,44	0,14	-0,09	-0,18	-0,28	-0,42	-0,03
二つの庭	0,58	-0,03	-0,17	-0,23	-0,21	0,02	0,01
道標	0,38	0,12	-0,06	-0,36	-0,07	0,29	-0,05

Table 3.2: The selected works by Miyamoto Yuriko listed according to the year in which were written. The right part of the table displays the weights of the writing corresponding to the 7 topics defined by the set of topic words listed above.

Chapter 4

Sentiment analysis of newspaper articles addressing the problem of Japanese officials' visits to Yasukuni shrine

This chapter addresses the problem of sentiment analysis, which represents a fruitful topic in the field of text mining. We show its potential utilization for Japanese studies on the example of sentiment analysis of both Japanese and foreign newspaper articles written in English.

4.1 Research problem and methods

The research topic of this chapter is the Yasukuni shrine controversy viewed from both the Japanese and foreign perspective. More specifically, we pose a question whether newspapers in South Korea and China tend to be more critical with respect to Japanese officials visits to Yasukuni shrine than Japanese or western newspapers, or vice versa. To this end, we employ the so-called sentiment classification, which is one of the methods used in the field of sentiment analysis.

The first part of this chapter reviews briefly basic facts about Yasukuni shrine and the history of controversial visits payed by Japanese officials including several prime ministers.

Then we move on to the introduction to sentiment analysis itself, with special focus on the sentiment classification. Next part refers to the methods we decided to use in our concrete case of Yasukuni shrine articles. Finally, we present the results obtained from the analysis and comment on the drawbacks of our approach and usability of the method of sentiment analysis in the case of this and similar problems.

4.2 Yasukuni Shrine

Yasukuni Shrine is a Shinto shrine that was established by the Emperor Meiji in 1869, originally named as Tokyo Shokon-sha [16]. Its was renamed as Yasukuni in 1879 which can be translated as "preserving peace for the entire nation". Enshrined and consequently considered a divinity of Yasukuni is everybody who sacrificed his or her life for the nation during both domestic and international wars since 1853, including the Second World War, leading to 2,466,000 divinities enshrined at Yasukuni Shrine today. All the divinities are worshipped at the same footing regardless of their social status, including 14 class A war criminals who were enshrined in 1978.

In the worshipping of these class A war criminals lies the roots of international disputes. The most prominent critiques are China and South Korea whose citizens suffered the atrocities inflicted by the Japanese army in the first half of the 20th century. They argue that Yasukuni represents a symbol of Japanese military past and helps to strengthen Japanese nationalism and militarism.

Nevertheless, what draws the attention of neighbouring countries most, are the visits paid by Japanese politicians. Even though some Japanese officials including Prime ministers paid their visits to Yasukuni Shrine before, it was Prime Minister Nakasone Yasuhiro who went there on the first official visit in 1985. This event drew a sharp criticism from both China and Korea, resulting in Nakasone's refraining from another official visits.

The issue of Yasukuni Shrine came to the fore again in 2001 when Prime Minister Koizumi Jun'ichiro announced he would visit Yasukuni while in office. Shinzo Abe visited Yasukuni Shrine serving as a Prime Minister as well. These visits paid by Japanese officials lead to further worsening of international relations with the neighbouring countries [16].

4.3 Sentiment analysis – methods

In this section, which rely heavily on [1], we elaborate on the topic of sentiment analysis also called opinion mining in the literature. The sentiment analysis works with several core terms. One of the most crucial concepts is that of the opinion target which is the object of the expressed opinion. On the other side stands the so-called opinion source simply referring to the holder of the given opinion. We are usually concerned with the polarity of the opinion, which can be positive, negative or neutral.

Another central terms are subjectivity and objectivity. Objective sentences are those statements that provide factual information. On the contrary, subjective sentences contain feelings and personal views. Obviously, while subjective sentences are

more likely to express opinions, objective ones usually refrain from doing so. However, by no means can we consider subjective sentences to be equivalent to those containing opinions and vice versa. Throughout this thesis we do not take the matter of subjectivity/ objectivity into consideration and assess the texts only from the aspect of the polarity of the opinions.

All in all, to obtain comprehensive information about the opinions expressed in the text, we need to identify the opinion source and the opinion target and classify the opinion polarity.

4.3.1 Sentiment classification

Let us now focus on the most prominent task in the sentiment analysis nowadays – the so-called sentiment classification [1]. It considers the whole document as one unit and classifies it as positive, negative or neutral. The sentiment classification problem is based on one assumption, namely that the opinions have the same opinion source and are directed to one opinion target. This assumption is usually satisfied in the case of product reviews and twitter comments, which have represented a fruitful areas for sentiment analysis recently.

In this thesis we consider lexicon based methods that rely on a dictionary of the so-called opinion words that are assigned a polarity score. Such a dictionary can be created using one of three leading approaches. First approach is to collect the opinion words and annotate them manually, which is on one hand very accurate, but extremely demanding on the other hand.

Another approach – dictionary based approach – exploits a small list of labeled seed opinion words and an online dictionary. Its task is to expand the existing list by searching for their synonyms and antonyms in the online dictionary. The opinion word list then grows with every iteration of this process during which new words are added. This approach allows for a large annotated dictionary, on the other hand it cannot be used in cases we aim at domain specific dictionaries (contrary to the corpus based approach).

Finally, let us comment on the corpus based opinion word lists. Similarly to the aforementioned dictionary based approach, it starts with a limited list of seed opinion words – adjectives. To expand this list, syntactic and logical rules are used. These assume that adjectives joined by conjunction ”and” (presenting non-constrasting ideas) have the same polarity, whereas contrasting conjunctions such as ”but”, ”however” signify different polarity of the adjectives. These rules are applied to a large corpus to determine the polarity of the adjectives used in the documents. One significant ad-

vantage of this approach over the dictionary based approach is its domain dependency stemming from the corpora we use.

4.4 Sentiment classification of newspaper articles – methods

In our analysis we resorted to lexicon-based approach to classify the newspaper articles. There exist several dictionaries containing the information about the polarity of the words. They vary in the number of words included. Among the most widely used dictionaries are the SentiWordNet [17] and SentiWords [18]. SentiWordNet considers the polarity of the given word as well as its objectivity. SentiWordNet groups the words into the so-called synsets, which are simply sets of synonyms. Every synset is in this case assigned both positivity (pos) and negativity (neg) score, each in the range [0,1]. The objectivity score of a synset is then computed according to the formula obj = 1 − pos − neg. The SentiWordNet contains approximately 155000 annotated English words .

In this thesis, however, we decided to use the SentiWords dictionary. This opinion words list is based on the polarity information derived from the SentiWordNet and can be readily used for our analysis.

An entry in SentiWords is in the format 'word#part-of-speech number from -1 to 1'. SentiWords recognizes four different part of speech: noun (n), adjective (a), verb (v) and adverb (r). The polarity score of a word then differs according to its concrete usage in a sentence taking into account the part of speech becomes necessary, as can be seen in the following example

visit#n 0

visit#v 0.37802.

We used the following algorithm to evaluate the overall polarity of our texts:

1. Read the file and do the preprocessing (including the part of speech tagging and lemmatization)

2. Take every token from the original text and try to mach it with the corresponding lemma in the SentiWords:

   - if such a token is not covered by the SentiWords, ignore it

   - if it corresponds to some entry in the SentiWords:

- if the token is not preceded by negation (not, less, no, never, nothing, nowhere, hardly, barely, scarcely, nobody, none) then find its polarity and write it down
- if the token is preceded by negation then multiply ist polarity by -1 and write down the result

3. Sum up all the sentiment polarities obtained in the procedure above and divide this number by the overall number of tokens to acquire the average polarity of one token in the given text.

4.5 Sentiment classification of newspaper articles – results

In this section we apply the above summarized methods to sentiment analysis of the newspaper articles about Yasukuni shrine. We decided to use digital databases available online, as these readily provide texts in the digital format which is the necessary premise for text mining.

We consider four newspapers publishing companies: The Guardian (https://www.theguardian.com/us) that represent the westerner perspective, the most popular Chinese newspapers China Daily (http://www.chinadaily.com.cn/), Korean The Chosun Ilbo (http://english.chosun.com/) and finally the oldest English written Japanese newspapers The Japan Times (https://www.japantimes.co.jp/). We restricted our analysis to articles addressing the problem of Yasukuni shrine solely in the relation to the Japanese officials (Prime Ministers and politicians) visits to this shrine. Moreover, to achieve better accuracy, we tried to identify articles discussing the same event. Tables 4.1, 4.2, 4.3 and 4.4 then refer to the headline and date of issue of the articles we use in the analysis.

After downloading the texts, we used the script (in Python) with the algorithm described in the previous section to derive the average polarity of the text. In this particular case of the Yasukuni shrine controversy, we were interested in the amount of negative sentiment included in the articles, therefore we considered only the opinion words with the negative sentiment. As a consequence, the results summarized in Table 4.5 show negative polarity (note that the value must be in the range [-1,0]).

In the second part of the analysis we focused solely on the newspapers headings, inspired by several previous studies such as [19] and researches cited thereof. In this case we searched for newspapers headlines according to their relevance to the topic, thus choosing those that are most likely to have caught the eyes of the readers when

typing "Yasukuni". We collected 20 headlines from each of the four newspapers and performed the same algorithm as in the case of the whole articles with one exception: we looked at the negative sentiment per headline (not per token). The results are shown in Table 4.6 (the resulting negativity score is not necessarily bound to be in the range [-1,0] since we do not normalize it by the number of tokens).

The results obtained from this analysis are not completely in accordance with our expectations. We anticipated China Daily and The Chosun Ilbo to be more critical since both China and South Korea suffered from the Japanese atrocities most during the Second World War and consequently belong to the most prominent critiques of Japanese attitude to the war past. At the same time, we expected The Japan Times would be most liberal. However, both the analysis of the articles and that of the headlines suggests that The Guardian and The Chosun Ilbo tend to use more negative language when referring to the Yasukuni visits while China Daily and The Japan Times seem to use less negative sentiment. The positive aspect of the outcome is the consistency of the results obtained from the sentiment analysis of the articles and the processing of the headlines. This undoubtedly contributes to the credibility of the conclusions, despite them not being in accordance with our hypothesis.

However, this analysis has several major drawbacks. Firstly, as can be inferred from the Table 4.5, the overall negativity score is by no means consistent thorough different articles published by the same company, making the results somewhat unreliable. Secondly, the assumption of one opinion source and the same opinion target is not always satisfied in the articles. That is, even though the opinions expressed are usually targeted at the person paying the visit, few of them present the visitors response as well.

As a result, more complex analysis would be necessary in order to eliminate those shortcomings and obtain more accurate results. To solve the first problem, we would need to collect a larger set of articles, preferably also from other publishers. The second deficiency of our approach could be amended by detecting those parts of the text that refer to the visitor while ignoring other statements. To this end, we would have to identify parts with direct and reported speeches, detect the opinion source and analyze only the relevant parts. However, such analysis requires named entity detection, which is beyond the scope of this thesis.

	Headline	date of issue
1.	Koizumi's final shrine trip draws protests	15. 8. 2006
2.	Japan shrine visit angers South Korea	22. 4. 2013
3.	Two Japanese ministers visit war shrine	15. 8. 2013
4.	China summons Japanese ambassador over war shrine visit	18. 10. 2013
5.	Japan's Shinzo Abe angers neighbours and US by visiting war dead shrine	26. 12. 2013
6.	China protests at Japanese PM' s latest WW2 shrine tribute	17. 10. 2014
7.	Anger as Japanese minister visits 'war crimes' shrine after Pearl Harbor trip	26. 12. 2016

Table 4.1: Articles from The Guardian.

	Headline	date of issue
1.	Koizumi's provocation condemned	16. 8. 2006
2.	China and South Korea criticize visits to shrine	23. 4. 2013
3.	China, S Korea condemn Japan over war shrine visit	16. 8. 2013
4.	About 160 Japanese lawmakers visit Yasukuni Shrine	18. 10. 2013
5.	China strongly condemns Abe's shrine visit	26. 12. 2013
6	3 Japanese female ministers visit notorious Yasukuni Shrine	18. 10. 2014
7.	Japan DM visits notorious Yasukuni Shrine	29. 12. 2016

Table 4.2: Articles from China Daily.

	Headline	date of issue
1.	Korea Slams Koizumi's Aug. 15 War Shrine Worship	15. 8. 2006
2.	Foreign Minister Cancels Japan Trip	23. 4. 2013
3.	Japanese Cabinet Ministers Visit Yasukuni War Shrine	16. 8. 2013
4.	Abe Makes Another Offering to Yasukuni Shrine	18. 10. 2013
5.	Abe Angers Neighbors with War Shrine Visit	27. 12. 2013
6.	Japan's Abe Sends Offering to War Shrine	18. 10. 2014
7.	Japan Criticized by East Asian Neighbors for Visit to Controversial WWII Shrine	30. 12. 2016

Table 4.3: Articles from The Chosun Ilbo.

	Headline	date of issue
1.	Defiant Koizumi visits Yasukuni	16. 8. 2006
2.	Record 168 lawmakers visit Yasukuni	24. 4. 2013
3.	Three ministers visit Yasukuni on surrender day anniversary; Abe refrains	15. 8. 2014
4.	Around 70 Japanese lawmakers visit war-linked Yasukuni Shrine for autumn festival	18. 10. 2013
5.	Abe visits Yasukuni, angering Beijing and Seoul	26. 12. 2013
6.	Abe sends ritual offering to Yasukuni; several lawmakers visit shrine	17. 10. 2014
7.	Defense chief Inada disrupts Abe's historic moment by visiting Yasukuni	29. 12. 2016

Table 4.4: Articles from The Japan Times.

Article No.	The Guardian	China Daily	The Chosun Ilbo	The Japan Times
1.	-0.0418	-0.0268	-0.0335	-0.0374
2.	-0.0405	-0.0261	-0.0256	-0.0296
3.	-0.0465	-0.0392	-0.0521	-0.0364
4.	-0.0459	-0.0289	-0.0382	-0.0249
5.	-0.0363	-0.0409	-0.0565	-0.0376
6.	-0.0437	-0.0364	-0.0483	-0.0201
7.	-0.0375	-0.0228	-0.0345	-0.0386
average	-0.0417	-0.0315	-0.0413	-0.0320

Table 4.5: The average negative opinion per token in the articles used in the analysis.

The Guardian	China Daily	The Chosun Ilbo	The Japan Times
-1.085	-0.258	-0.489	-0.256

Table 4.6: The negative opinion per headline.

Chapter 5

Automatic summarization of academic articles

This chapter focuses on the automatic summarization techniques and their practical utilization for researchers who need to process a large amount of academic texts.

5.1 Research problem and methods

The aim of this chapter is to illustrate the method of text summarization on both English and Japanese sample academic papers. We chose two types of texts: technically oriented ones (text mining) and articles from humanities (Japanese studies).

The central question we pose in this chapter is whether text summarization techniques could provide viable tools for researchers in Japanese studies or not. That is, are the summarization methods able to generate relevant and accurate summaries for a wide range of academic articles related to Japanese studies? If so, this method could possibly be of use for Japanese studies researchers so that they could extract relevant and important information from large amount of texts with minimum effort.

In the field of Japanese studies we usually work with English as well as Japanese primary and secondary literature. Moreover, these sources vary considerably in the style and vocabulary they use ranging from relatively technical texts (linguistics) to texts relying on elaborate language structures (more humanitely oriented texts). In order to probe whether automatic summarization depends on the text domain, we study two English and two Japanese texts considering technically oriented texts as well as humanitely oriented texts in both cases.

Most of the summarization techniques focus on summarization of big corpora. However, such a task is far beyond the scope of this thesis; therefore we resort to the so-called graph method called TextRank algorithm (for details we refer to (5.2.2))

which is applicable in the case of a single-document summarization. To evaluate the performance of this summarization algorithm, we compare these automatic summaries with manually created summaries afterwards.

We review the basics of automatic summarization in the first part of this chapter. Then we move on to the practical demonstration of the TextRank algorithm on the four sample texts and evaluate the quality of the summaries. We conclude this chapter with a discussion about the applicability of this method to Japanese studies.

5.2 Text summarization – methods

A summary is defined as a text produced from one or more texts that conveys important information of the original text, the length of which does not exceed the length of the original text [7].

In this section we restrict ourselves to extractive summarization methods, since these usually outperform the abstractive ones and are commonly used nowadays [1], [20]. As was already mentioned in Chapter 1, extractive summarization identifies the most significant sentences in a document or a set of documents and produces a summary that consists of these sentences ordered as in the original text.

We distinguish between two basic types of automatic summarization approaches, namely to so-called topic representation approach and indicator representation approach. In the case of the topic representation, the algorithm aims at identifying of the topic described in the text. In the latter case, the indicator representation assigns a list of indicators to every sentence. These are characteristics such as the length or position in the document.

The extractive summarization process can be divided into three subsequent steps:

- creation of intermediate representation
- scoring of the sentences
- selecting the sentences that would constitute the summary

Intermediate representation is a form of the original text which presents core information about the text and enables us to determine the importance of the respective sentences. It is dependent on the type of technique we want to perform. For instance, in the case of frequency-based algorithms, the intermediate representation is a list of words with the corresponding weights.

5.2.1 Topic representation techniques

Let us now focus on the topic representation techniques utilized in automatic summarization. Among the most widely used techniques are topic words approach, frequency-driven approaches and Latent semantic analysis.

Algorithms based on the topic words identification rely on detection of the so-called topic signatures which are words that describe well the topic of the text. These are words that occur often in the given text while being rare or absent in other texts. As a result, this method usually requires a big corpus. Each sentence is then assigned a score according to the number of topic signatures within it or the ratio of topic signatures and other words in the sentence.

Frequency-driven approaches work with more elaborate process of identification of the words describing the topic. The basic model is based on the word probability, which is simply a number of occurrences of the given word divided by the overall number of the words in the document. This basic model can be modified when we consider evaluating the importance of the words using the TF-IDF weighting (for details we refer to Section 3.2.1), since it enables us to disregard words that do not add meaningful content (such as stop-words). Another technique takes advantage of the Latent semantic analysis (see Section 3.2.2).

5.2.2 Indicator representation techniques

In this section we comment on the problem of indicator representation techniques, more specifically, we often deal with graph methods.

One of the most popular graph method is the TextRank algorithm which is based on the PageRank algorithm [24]. The intermediate representation of the input text is in the form of a graph, where sentences are represented by vertices and the edges between them are labeled by weights corresponding to how similar the two sentences are (larger weights for more similar sentences) . The so-called cosine similarity is often used as a measure of the similarity while the weights of the words are usually determined by the TF-IDF weighting. The sentence is more likely to be selected for the summary when connected to many other sentences. This method can be readily applied to any language since it only works with respective words in the sentences.

5.3 Automatic summarization of articles – results

For the summarization of English articles the following two texts were selected:

a) linguistic article: Preprocessing Techniques for Text Mining [4] (abbreviated)

b) humanities oriented article: The Meiji Restoration and Modernization [21]

In the case of Japanese-written articles the summarization was performed on two articles:

a) linguistic article: 日本におけるテキストマイニングの応用 [22] (abbreviated)

b) humanities oriented article: 明治維新：近代国家への歩み [23].

We have at disposal a summary and a list of keywords created manually by people proficient in these languages with an academic background. The texts with manually and automatically selected sentences are presented in Appendix.

For the summarization itself we use predefined graph method implemented in Python. As mentioned in Section 5.2.2, the only indispensable information for the TextRank algorithm is that about the word boundaries. Therefore, to adjust this method for Japanese, we need to perform tokenization first.

To enhance the perfomance of the summarizer, we resolved to apply more complex preprocessing. In the case of the English texts, we perfomed lemmatization of the tokens and subsequently removed the stop-words according to a stop-word list predefined in NLTK.

For the Japanese texts we applied the above procedure with slight modification. Since a list of stop-words for Japanese that would be predefined in MeCab does not exists, we filtered out all tokens with the exception of nouns and verbs.

Note that the number of sentences selected for the automatic summaries corresponds to the number of sentences in the manual summaries, which subsequently facilitates the statistical analysis.

Let us first focus on the list of keywords which are displayed in Table 5.2 and 5.1. These tables contain 10 keywords for both of the texts generated manually as well as automatically.

A brief observation of these keywords reveals a considerable difference between those created manually and those generated automatically. This discrepancy is more prominent in the case of the Japanese texts. One of the reasons is undoubtedly the fact, that the algorithm is not able to detect collocations and compound words which comprise the majority of the manually selected keywords of the Japanese articles. The ratio of coincidence is better in the case of English articles, amounting to the ratio 60% for article a) and 40% for article b), respectively.

Regarding the relevance of the automatically generated keywords, with the exception of the Japanese text b), they capture the central terms in the texts even though

they do not always coincide with the manually selected ones.

Secondly, we evaluate and compare the manual and automatic summaries themselves employing basic approaches of statistics. The characteristics of the input texts as well as the results of the analysis are displayed in Table 5.3. The approach that we follow can be viewed as a modification of the evaluation based on the so-called F-measure [25].

The F-measure computes characteristics of the human-written summary and the automatic summary. Those are the recall and the precision that are defined as (# stands for "number of")

$$recall = \frac{\#\text{ human selected sentences} \cap \#\text{ automatically selected sentences}}{\#\text{ human selected sentences}}$$
$$precision = \frac{\#\text{ human selected sentences} \cap \#\text{ automatically selected sentences}}{\#\text{ automatically selected sentences}}$$

.

The accuracy of a summary is then expressed by the F-measure reading

$$F - measure = 2\frac{recall * precision}{recall + precision}. \tag{5.1}$$

The F-measure which ranges from 0 to 1 represents an elementary method to evaluate the accuracy of a summary. Nevertheless, it is perfectly applicable to our analysis, since we possess the human constructed summary consisting of identical sentences to those from the original text. Moreover, since we requested automatic summaries of the same length as those human-written summaries, the above formula reduces to

$$F - measure = recall = precision. \tag{5.2}$$

In the following we use this terminology to evaluate our summaries.

Crucial feature of the summary is its length. By the length of the summary we understand the ratio of the number of sentences comprising the summary to the overall number of sentences in the given text. This ratio can also be interpreted as the probability of a single sentence to be selected for the summary. In Table 5.3 we observe that this ratio ranges from 0.29 to 0.43 for the four texts considered.

The central step in the analysis relies on the comparison of the manual (the so-called reference summary) with the automatic summaries. This is reduced to a simple task of evaluating the number of sentences that coincide in the reference and automatic

summaries. The ratio of this number to the number of sentences in the summary then corresponds to the F-measure. The F-measure then represents the crucial indicator to what extent the automatic summarization methods are reliable under the assumption that we consider the human-made summaries completely relevant and accurate. If we accept this supposition, we aim at 100% correspondence of the sentence selection or, in other words at F-measure = 1.

However, the analysis reveals considerable discrepancy between this ideal case and the real data we obtained. The F-measure ranges from mere 20% to 51%. In order to interpret this result and assess the performance of the automatic summarizer in a meaningful way, we compare the F-measure value with the probability of coincidence in case we select the sentences randomly. As mentioned above, the probability of a single sentence to appear in the summary is the ratio of the number of sentences selected to their overall number. If we assume the two summaries (automatic and reference) completely unrelated then the probability of the co-occurrence of a single sentence in these two summaries is the ratio to the power of two. Based on this we can calculate the F-measure for random summaries of the same length as that of our reference summaries (this F-measure is displayed in Table 5.3 under the label F-measure – random selection). Taking this characteristics into consideration, we see that the automatic summarization method was able to capture certain patterns in the English text a) and b) and Japanese text b), where it outperformed the random selection. Its performance proved to be best in the case of the article The Meiji Restoration and Modernization, where the ratio of the F-measure of summaries to the F-measure for random selection was highest.

It remains an open question whether the automatic summarization techniques represent a viable method in Japanese studies. The above analysis yielded results that can only be regarded as preliminary ones. In a more comprehensive analysis several aspects have to be considered. The fundamental problem is the number of texts; a reliable statistics requires a large number of samples.

Another shortcoming lies in the evaluation method used in our analysis. We relied on the F-measure that compares sentences selected in the reference summary to those chosen automatically. However, several sentences in the text may convey the same information in a different way, making the selection highly subjective for humans. In [26] the author pointed out that the overlap of scientific extracts created by 6 different people was as low as 8%. Moreover, another results suggest that the recall can vary from 25% to 50% for extractive summaries created by different humans. Due to the aforementioned arguments, F-measure cannot be considered a reliable technique in the

assessment of summaries. We suggest using more refined methods of evaluation such as the ROUGE technique [25] in future research.

Manual selection	Automatic selection	Manual selection	Automatic selection
word	text	restoration	Japan
text	character	Japan	Japanese
process	word	emperor	nation
token	token	independence	western
document	process	reform	political
stem	use	colony	government
language	document	government	samurai
content	data	military	military
classification	format	war	industry
retrieve	language	nation	period

Table 5.1: English texts: 10 keywords selected both manually and automatically for article a) (on the left) and article b) (on the right).

Manual selection	Automatic selection	Manual selection	Automatic selection
テキストマイニング	形態素	不平等条項	アジア
相関関係	ソフトウェ	尊王攘夷	綿織物
定量分析	テキスト	薩長同盟	無条件
応用	インパクト	王政復古の大号令	不平等
日本語	MeCab	戊辰戦争	江戸城
形態素	マイニング	五箇条の御誓文	認める
分割	データ	廃藩置県	抑える
係り受け関係	パッケージ	天皇	代わる
共起関係	組み込む	自由民権運動	戊辰戦争
統計的手法	Juman	大日本帝国憲法	クーデター

Table 5.2: Japanese texts: 10 keywords selected both manually and automatically for article a) (on the left) and article b) (on the right).

	English		Japanese	
Article	a)	b)	a)	b)
Overall number of sentences	47	94	43	77
Number of sentences in the summary	12	27	15	33
Ratio of sentences selected to their overall number	0.26	0.29	0.35	0.43
Number of sentences coinciding	4	13	3	17
F-measure – random selection	27%	28%	34%	42%
F-measure – summaries	33%	48%	20%	51%

Table 5.3: Automatic summarization of two Japanese and two English articles denoted as a) and b): number of sentences in the articles, lengths of the summaries, the corresponding F-measure of the automatic summaries and F-measure for randomly selected summaries.

Conclusion

This thesis was devoted to the analysis of possible utilization of certain text mining techniques in Japanese studies. The review of basic text mining methods was provided in the first chapter and more detailed discussion of one important step – preprocessing – was presented in the second chapter.

To test some of the methods in Chapter 3 we applied clustering to selected works written by two Japanese proletarian writers. The writings were automatically classified into groups according to topic similarity with the help of Latent semantic analysis and the results were compared with existing literary studies. The comparison yielded very good coincidence pointing to the suggestion that this method can indeed be of use in certain parts of Japanese studies to sort large corpora of texts.

Chapter 4 addressed the problem of Japanese officials' visits to Yasukuni shrine viewed from the perspective of both foreign and Japanese newspapers. The aim of this research topic was to evaluate negative sentiment expressed in British, Chinese, Korean and Japanese newspaper articles and their headlines. The adaptability of the method of sentiment analysis to Japanese studies stays arguable. The approach used in this thesis is rather oversimplified, since we did not consistently identify neither the opinion source nor the opinion target. More refined analysis is required in order to assess the viability of the sentiment mining methods in Japanese studies.

Techniques of automatic summarization and their application to English and Japanese texts were the central point of Chapter 3. To evaluate the reliability of the TextRank algorithm, we compared the automatic summaries with human-made reference summaries. The algorithm performed better than simple random selection of sentences only in two of all four articles and the highest number of coincidence achieved in comparison with the reference summaries amounted to mere 51%. We believe that the reliability of automatic summaries would score better if we used more suitable technique of evaluation such as the ROUGE method or had more statistically relevant data supported by higher number of both articles and manually made summaries.

Text mining methods as a whole definitely present a potential merit in Japanese studies. However, in order to obtain relevant and reliable results, it is crucial to fine-

tune the methods for specific tasks, which requires much deeper and thorough analysis than the one described in this thesis. This is probably true to any other usage of text mining since precise fine-tuning is needed in most machine learning methods. For the purpose of future research we recommend narrowing down the research topic to one specific problem and focusing on a wide range of methods provided by text mining. Such approach would probably identify the optimal technique for the given problem and obtain much more conclusive results.

Bibliography

[1] Aggarwal, Charu C., and Zhai, ChengXiang. Mining Text Data. Springer, 2012.

[2] Weiss, Sholom M., Indurkhya, Nitin, and Zhang Tong. Fundamentals of Predictive Text Mining. Springer, 2010.

[3] Solka, Jeffrey L. "Text Data Mining: Theory and Methods". Statistics Survey, vol. 2, 2008, pp. 94-112.

[4] Gurusamy, Vairaprakash, and Kannan, Subbu. "Preprocessing Techniques for Text Mining". Recent Trends and Research Issues in Computer Science: Proceeding of the conference Recent Trends and Research Issues in Computer Science, Podi, 2014

[5] Manning, Christopher G., Raghavan, Prabhakar, and Schütze, Hinrich. Introduction to Information Retrieval. Cambridge University Press, 2008.

[6] Sethunya Joseph R., Hlomani, Hlomani, Letsholo, Keletso, Kaniwa, Freeson, and Sedimo, Kutlwano. "Natural Language Processing: A Review". International Journal of Research in Engineering and Applied Sciences, vol. 6, 2016, pp. 207-210.

[7] Allahyari, Mehdi, Pouriyeh, Seyedamin, Assefi, Mehdi, Safaei, Saeid, Trippe, Elizabeth, Gutierrez, Juan, and Kochut, Krys. "Text Summarization Techniques: A Brief Survey". International Journal of Advanced Computer Science and Applications, vol. 8, 2017, pp. 397-405.

[8] Allahyari, Mehdi, Pouriyeh, Seyed A., Assefi, Mehdi, Safaei, Saied, Trippe, Elizabeth D., Gutierrez, Juan B., and Kochut, Krys J.. "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques" , 2017. URL https://arxiv.org/pdf/1707.02919.pdf

[9] Patel, Sangita N., and Choksi, Jignya B. "A Survey of Sentiment Classsifiction Techniques." Journal 4 Research, vol 1, 2015, pp. 15-20.

[10] Catalinak, Amy. "Quantitative Text Analysis with Asian Languages: Some Problems and Solutions." Polimetrics, vol. 1, 2014, pp. 14-17.

[11] Den, Yasuharu, Nakamura, Junpei, Ogiso, Toshinobu, and Ogura, Hideki. "A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation." International Conference on Language Resources and Evaluation: Proceeding of the Conference International Conference on Language Resources and Evaluation, Morroco, 2008.

[12] Kudo, Taku, Yamamoto, Kaoru, and Matsumoto, Yuji . "Applying conditional random fields to Japanese morpho-logical analysis." Empirical Methods in Natural Language Processing: Proceeding of the Conference on Empirical Methods in Natural Language Processing, Barcelona, 2004.

[13] Novák, Miroslav. Japonská literatura 2., Státní pedagogické nakladatelství Praha, 1989.

[14] Kinjo, Gillian. "Rivers of the soul by Miyamoto Yuriko : biography of the author, translation, discussion." MA Thesis. University of Canterbury, 1984.

[15] Mostow, Joshua , Denton, Kirk, Fulton, Bruce, and Orbaugh, Sharalyn. The Columbia Companion to Modern East Asian Literature, Columbia University Press, 2003

[16] Michiaki, Okuyama. "The Yasukuni Shrine Problem in the East Asian Context: Religion and Politics in Modern Japan." Politics and Religion Journal, vol 3, 2009, pp. 235-251.

[17] Esuli, Andrea, and Sebastiani, Fabrizio. "SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining". Conference on Language Resources and Evaluation: Proceeding of conference Conference on Language Resources and Evaluation, Italy, 2006.

[18] Gatti, Lorenzo, Guerini, Marco, Turchi, Marco. "SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis", 2015. URL https://arxiv.org/pdf/1510.09079.pdf

[19] Somanath, Chavan S., and Yash, Chavan. "Sentiment Classification of News Headlines on India in the US Newspaper: Semantic Orientation Approach vs Machine Learning", 2019. DOI: 10.13140/RG.2.2.34008.75522

[20] Das, Dipanjan, and Martins, André. "A survey on automatic text summarization.", 2007. URL https://www.cs.cmu.edu/ nasmith/LS2/das-martins.07.pdf

[21] Asia for Education. The Meiji Restoration and Modernization. URL http://afe.easia.columbia.edu/special/japan_1750_meiji.htm

[22] 齋藤, 朗宏. ”日本におけるテキストマイニングの応用”. URL https://www.kitakyu-u.ac.jp/economy/study/pdf/2011/2011_11.pdf

[23] 河合, 敦. ”明治維新：近代国家への歩み”. URL https://www.nippon.com/ja/views/b06902/

[24] Barrios, Federico, L´opez, Federico, Argerich, Luis, and Wachenchauzer, Rosita. ”Variations of the Similarity Function of TextRank for Automated Summarization”, 2016. URL http://arxiv.org/abs/1602.03606

[25] Lloret, Elena, Plaza, Laura, and Aker, Ahmet. ”The challenging task of summary evaluation: an overview”, Language Resources and Evaluation, vol 52, 2017, pp. 101-148.

[26] Nenkova, Ani. ”Summarization evaluation for text and speech: issues and approaches”, Ninth International Conference on Spoken Language Processing: Proceeding of conference Ninth International Conference on Spoken Language Processing, USA, 2006.

Appendix

We append the texts used in Chapter 5 for reference. Green color highlights the sentences that appear in both the manual and automatic summary. Red color marks the sentences present in the human-made summaries and absent in the automatic summary. On the contrary, blue color highlights sentences selected only in the automatic summary.

English text a)

Preprocessing Techniques for Text Mining [4]

Introduction

Text pre-processing is an essential part of any NLP system, since the characters, words, and sentences identified at this stage are the fundamental units passed to all further processing stages, from analysis and tagging components, such as morphological analyzers and part-of-speech taggers, through applications, such as information retrieval and machine translation systems.

It is a collection of activities in which Text Documents are pre-processed. Because the text data often contains some special formats like number formats, date formats and the most common words that unlikely to help Text mining such as prepositions, articles, and pro-nouns can be eliminated.

Need of Text Preprocessing in NLP System

To reduce indexing (or data) file size of the Text documents. Stop words accounts 20-30% of total word counts in a particular text documents. Stemming may reduce indexing size as much as 40-50%. To improve the efficiency and effectiveness of the IR system. Stop words are not useful for searching or Text mining and they may confuse the retrieval system. Stemming used for matching the similar words in a text document.

Tokenization

Tokenization is the process of breaking a stream of text into words, phrases, symbols, or other meaningful elements called tokens . The aim of the tokenization is the exploration of the words in a sentence. The list of tokens becomes input for further processing such as parsing or text mining. Tokenization is useful both in linguistics (where it is a form of text segmentation), and in computer science, where it forms part of lexical analysis. Textual data is only a block of characters at the beginning. All processes in information retrieval require the words of the data set. Hence, the requirement for a parser is a tokenization of documents. This may sound trivial as the text is already stored in machine-readable formats. Nevertheless, some problems are still left, like the removal of punctuation marks. Other characters like brackets, hyphens, etc require processing as well. Furthermore, tokenizer can cater for consistency in the documents. The main use of tokenization is identifying the meaningful keywords. The inconsistency can be different number and time formats. Another problem are abbreviations and acronyms which have to be transformed into a standard form.

Challenges in Tokenization

Challenges in tokenization depend on the type of language. Languages such as English and French are referred to as space-delimited as most of the words are separated from each other by white spaces. Languages such as Chinese and Thai are referred to as unsegmented as words do not have clear boundaries. Tokenizing unsegmented language sentences requires additional lexical and morphological information. Tokenization is also affected by writing system and the typographical structure of the words.

Structure of languages can be grouped into three categories:

Isolating: Words do not divide into smaller units. Example: Mandarin Chinese.

Agglutinative: Words divide into smaller units. Example: Japanese, Tamil.

Inflectional: Boundaries between morphemes are not clear and ambiguous in terms of grammatical meaning. Example: Latin.

Stop Word Removal

Many words in documents recur very frequently but are essentially meaningless as they are used to join words together in a sentence. It is commonly understood that stop words do not contribute to the context or content of textual documents. Due to their high frequency of occurrence, their presence in text mining presents an obstacle in understanding the content of the documents.

Stop words are very frequently used common words like ‘and’ , ‘are’ , ‘this’ etc. They are not useful in classification of documents. So they must be removed. However,

the development of such stop words list is difficult and inconsistent between textual sources. This process also reduces the text data and improves the system performance. Every text document deals with these words which are not necessary for text mining applications.

Stemming

Stemming is the process of conflating the variant forms of a word into a common representation, the stem. For example, the words: “presentation” , “presented” , “presenting” could all be reduced to a common representation “present” . This is a widely used procedure in text processing for information retrieval (IR) based on the assumption that posing a query with the term presenting implies an interest in documents containing the words presentation and presented.

English text b)

The Meiji Restoration and Modernization [21]

In 1868 the Tokugawa shôgun (”great general”), who ruled Japan in the feudal period, lost his power and the emperor was restored to the supreme position. The emperor took the name Meiji (”enlightened rule”) as his reign name; this event was known as the Meiji Restoration.

The Reign of the Meiji Emperor

When the Meiji emperor was restored as head of Japan in 1868, the nation was a militarily weak country, was primarily agricultural, and had little technological development. It was controlled by hundreds of semi-independent feudal lords. The Western powers —Europe and the United States —had forced Japan to sign treaties that limited its control over its own foreign trade and required that crimes concerning foreigners in Japan be tried not in Japanese but in Western courts. When the Meiji period ended, with the death of the emperor in 1912, Japan had a highly centralized, bureaucratic government; a constitution establishing an elected parliament; a well-developed transport and communication system; a highly educated population free of feudal class restrictions; an established and rapidly growing industrial sector based on the latest technology; and a powerful army and navy.

Japan had regained complete control of its foreign trade and legal system, and, by fighting and winning two wars (one of them against a major European power, Russia), it had established full independence and equality in international affairs. In a little

more than a generation, Japan had exceeded its goals, and in the process had changed its whole society. Japan's success in modernization has created great interest in why and how it was able to adopt Western political, social, and economic institutions in so short a time.

One answer is found in the Meiji Restoration itself. This political revolution "restored" the emperor to power, but he did not rule directly. He was expected to accept the advice of the group that had overthrown the shôgun, and it was from this group that a small number of ambitious, able, and patriotic young men from the lower ranks of the samurai emerged to take control and establish the new political system. At first, their only strength was that the emperor accepted their advice and several powerful feudal domains provided military support. They moved quickly, however, to build their own military and economic control. By July 1869 the feudal lords had been requested to give up their domains, and in 1871 these domains were abolished and transformed into prefectures of a unified central state.

The feudal lords and the samurai class were offered a yearly stipend, which was later changed to a one-time payment in government bonds. The samurai lost their class privileges, when the government declared all classes to be equal. By 1876 the government banned the wearing of the samurai's swords; the former samurai cut off their top knots in favor of Western-style haircuts and took up jobs in business and the professions.

The armies of each domain were disbanded, and a national army based on universal conscription was created in 1872, requiring three years' military service from all men, samurai and commoner alike. A national land tax system was established that required payment in money instead of rice, which allowed the government to stabilize the national budget. This gave the government money to spend to build up the strength of the nation.

Resistance and Rebellion Defeated

Although these changes were made in the name of the emperor and national defense, the loss of privileges brought some resentment and rebellion. When the top leadership left to travel in Europe and the United States to study Western ways in 1872, conservative groups argued that Japan should reply to Korean's refusal to revise a centuries old treaty with an invasion. This would help patriotic samurai to regain their importance. But the new leaders quickly returned from Europe and reestablished their control, arguing that Japan should concentrate on its own modernization and not engage in such foreign adventures.

For the next twenty years, in the 1870s and 1880s, the top priority remained do-

mestic reform aimed at changing Japan's social and economic institutions along the lines of the model provided by the powerful Western nations. The final blow to conservative samurai came in the 1877 Satsuma rebellion, when the government's newly drafted army, trained in European infantry techniques and armed with modern Western guns, defeated the last resistance of the traditional samurai warriors. With the exception of these few samurai outbreaks, Japan's domestic transformation proceeded with remarkable speed, energy, and the cooperation of the people. This phenomenon is one of the major characteristics of Japan's modern history.

Ideology

In an effort to unite the Japanese nation in response to the Western challenge, the Meiji leaders created a civic ideology centered around the emperor. Although the emperor wielded no political power, he had long been viewed as a symbol of Japanese culture and historical continuity. He was the head of the Shintô religion, Japan's native religion. Among other beliefs, Shintô holds that the emperor is descended from the sun goddess and the gods who created Japan and therefore is semidivine. Westerners of that time knew him primarily as a ceremonial figure. The Meiji reformers brought the emperor and Shintô to national prominence, replacing Buddhism as the national religion, for political and ideological reasons. By associating Shintô with the imperial line, which reached back into legendary times, Japan had not only the oldest ruling house in the world, but a powerful symbol of age-old national unity.

The people seldom saw the emperor, yet they were to carry out his orders without question, in honor to him and to the unity of the Japanese people, which he represented. In fact, the emperor did not rule. It was his "advisers," the small group of men who exercised political control, that devised and carried out the reform program in the name of the emperor.

Social and Economic Changes

The abolition of feudalism made possible tremendous social and political changes. Millions of people were suddenly free to choose their occupation and move about without restrictions. By providing a new environment of political and financial security, the government made possible investment in new industries and technologies.

The government led the way in this, building railway and shipping lines, telegraph and telephone systems, three shipyards, ten mines, five munitions works, and fifty-three consumer industries (making sugar, glass, textiles, cement, chemicals, and other important products). This was very expensive, however, and strained government finances, so in 1880 the government decided to sell most of these industries to private investors, thereafter encouraging such activity through subsidies and other incentives.

Some of the samurai and merchants who built these industries established major corporate conglomerates called zaibatsu, which controlled much of Japan's modern industrial sector.

The government also introduced a national educational system and a constitution, creating an elected parliament called the Diet. They did this to provide a good environment for national growth, win the respect of the Westerners, and build support for the modern state. In the Tokugawa period, popular education had spread rapidly, and in 1872 the government established a national system to educate the entire population. By the end of the Meiji period, almost everyone attended the free public schools for at least six years. The government closely controlled the schools, making sure that in addition to skills like mathematics and reading, all students studied "moral training," which stressed the importance of their duty to the emperor, the country and their families.

The 1889 constitution was "given" to the people by the emperor, and only he (or his advisers) could change it. A parliament was elected beginning in 1890, but only the wealthiest one percent of the population could vote in elections. In 1925 this was changed to allow all men (but not yet women) to vote.

To win the recognition of the Western powers and convince them to change the unequal treaties the Japanese had been forced to sign in the 1850s, Japan changed its entire legal system, adopting a new criminal and civil code modeled after those of France and Germany. The Western nations finally agreed to revise the treaties in 1894, acknowledging Japan as an equal in principle, although not in international power.

The International Climate: Colonialism and Expansion

In 1894 Japan fought a war against China over its interest in Korea, which China claimed as a vassal state. The Korean peninsula is the closest part of Asia to Japan, less than 100 miles by sea, and the Japanese were worried that the Russians might gain control of that weak nation. Japan won the war and gained control over Korea and gained Taiwan as a colony. Japan's sudden, decisive victory over China surprised the world and worried some European powers.

At this time the European nations were beginning to claim special rights in China —the French, with their colony in Indochina (today's Vietnam, Laos, and Cambodia), were involved in South China; the British also claimed special rights in South China, near Hong Kong, and later the whole Yangtze valley; and the Russians, who were building a railway through Siberia and Manchuria, were interested in North China. After Japan's victory over China, Japan signed a treaty with China which gave Japan special rights on China's Liaotung peninsula, in addition to the control of Taiwan.

But Japan's victory was short lived. Within a week, France, Russia, and Germany combined to pressure Japan to give up rights on the Liaotung peninsula. Each of these nations then began to force China to give it ports, naval bases, and special economic rights, with Russia taking the same Liaotung peninsula that Japan had been forced to return.

The Japanese government was angered by this incident and drew the lesson that for Japan to maintain its independence and receive equal treatment in international affairs, it was necessary to strengthen its military even further. By 1904, when the Russians were again threatening to establish control over Korea, Japan was much stronger. It declared war on Russia and, using all its strength, won victory in 1905 (beginning with a surprise naval attack on Port Arthur, which gained for Japan the control of the China Sea). Japan thus achieved dominance over Korea and established itself a colonial power in East Asia.

The Period 1912-1941

The Meiji reforms brought great changes both within Japan and in Japan's place in world affairs. Japan strengthened itself enough to remain a sovereign nation in the face of Western colonizing powers and indeed became a colonizing power itself. During the Taishô period (1912-1926), Japanese citizens began to ask for more voice in the government and for more social freedoms. During this time, Japanese society and the Japanese political system were significantly more open than they were either before or after. The period has often been called the period of "Taishô democracy." One explanation is that, until World War I, Japan enjoyed record breaking economic prosperity. The Japanese people had more money to spend, more leisure, and better education, supplemented by the development of mass media. Increasingly they lived in cities where they came into contact with influences from abroad and where the traditional authority of the extended family was less influential. Industrialization in itself undermined traditional values, emphasizing instead efficiency, independence, materialism, and individualism. During these years Japan saw the emergence of a "mass society" very similar to the "Roaring 20s" in the United States. During these years also, the Japanese people began to demand universal manhood suffrage which they won in 1925. Political parties increased their influence, becoming powerful enough to appoint their own prime ministers between 1918 and 1931.

At the end of World War I, however, Japan entered a severe economic depression. The bright, optimistic atmosphere of the Taishô period gradually disappeared. Political party government was marred by corruption. The government and military, consequently, grew stronger, the parliament weaker. The advanced industrial sector

became increasingly controlled by a few giant businesses, the zaibatsu. Moreover, Japan's international relations were disrupted by trade tensions and by growing international disapproval of Japan's activities in China. But success in competing with the European powers in East Asia strengthened the idea that Japan could, and should, further expand its influence on the Asian mainland by military force.

Japan's need for natural resources and the repeated rebuffs from the West to Japan's attempts to expand its power in Asia paved the way for militarists to rise to power. Insecurity in international relations allowed a right-wing militaristic faction to control first foreign, then domestic, policy. With the military greatly influencing the government, Japan began an aggressive military campaign throughout Asia, and then, in 1941, bombed Pearl Harbor.

Summary

The most important feature of the Meiji period was Japan's struggle for recognition of its considerable achievement and for equality with Western nations. Japan was highly successful in organizing an industrial, capitalist state on Western models. But when Japan also began to apply the lessons it learned from European imperialism, the West reacted negatively. redIn a sense Japan's chief handicap was that it entered into the Western dominated world order at a late stage. Colonialism and the racist ideology that accompanied it, were too entrenched in Western countries to allow an "upstart," nonwhite nation to enter the race for natural resources and markets as an equal. Many of the misunderstandings between the West and Japan stemmed from Japan's sense of alienation from the West, which seemed to use a different standard in dealing with European nations than it did with a rising Asian power like Japan.

Japanese text a)

日本におけるテキストマイニングの応用 [22]

言葉の分析においては、近年、テキストマイニングと呼ばれる研究分野が発展している。テキストマイニングとは、膨大なテキスト（文書）情報の中から有用な情報を掘り出す（マイニング）ことで、定型化されていないテキストデータを、一定のルールに従って定型化して整理し、データマイニングの手法を用いながら、相関関係などの定量分析を行う手法である。

文章の分析そのものには長い歴史がある。金 (2009b) によれば、19 世紀末には既に単語の長さの分布を用いた分析が行われている。自然言語テキストからの情報抽出についても、有村 (2003) によれば 1980 年代後半から研究されている。しか

し，データマイニングの一手法としてのテキストマイニングという名が与えられ、特に実用化が進んできたのは、インターネットや PC の普及に伴い電子化テキストが急激に増加し始めた 1990 年代後半になってから (那須川、2009) である。

ただ、その初期の研究は、主に理論研究と実用化のためのソフトウェアの発表が中心であり、応用研究は多くない。少数ではあるが見られた応用研究にしても、後述の那須川 (2001) を代表に、自分で必要なソフトウェアを開発するという方法を取られることが多かった。データマイニングの諸分野の中でも、応用が遅れているのは、テキストマイニングはある特定の言語への対応が求められるため、ある言語のために開発されたソフトウェアをそのまま他の言語に対して用いることができないという事情があったものと考えられる。最近では、樋口 (2004) の KHcoder を嚆矢とし、松村・三浦 (2009) の TinyTextMiner、金 (2009a) の MLTP に見られるように、日本語を分析することのできるフリーのソフトウェアも豊富である。形態素解析ソフトウェア MeCab を R に組み込んだパッケージ RMeCab(石田、2008) により、フリーの統計ソフトウェア R 上でもテキストマイニングは実行可能となっている。こういった背景から、最近では統計、テキストデータ分析の専門家ではない研究者による応用事例が数多く見られ、より一層の発展が期待されている。ただ一方で、那須川 (2009) が指摘するように、テキストマイニングという言葉やツールの普及と比べ、大きなインパクトにつながっている活用成功事例は少ない。Hearst (1999) の指摘する、貴重で新奇な情報を得てこそ真のテキストマイニングという立場からは、真のテキストマイニングに到達できていないとも言える。そこで、本論文では、日本におけるテキストマイニングの応用の現状を確認し、今後の発展の可能性について考察する。

テキストマイニングの技術

テキストマイニングの応用について考えるにあたり、テキストマイニングの基本的な考え方について解説する。テキストマイニングの入門書は数多く出ており、解説論文も少なくない。中でも、松村 (2008) が全体の流れを理解するにはわかりやすいので、解説は同論文を基本として行う。

形態素解析

日本語のように単語間の区切りが明示されていない言語は、分析に先立って文章を分かち書きし，形態素に分割する。形態素とは、「言語学で、意味を持った最小の音型。ヤマ（山）のように形態素一つで単語が構成される場合もあれば、ヤマカゼ（山風）のように複数の形態素が単語を構成する場合もある（大辞泉）」とされる。文章から形態素を探し出し、その形態素単位に分割することを形態素解析と呼ぶ。日本語の形態素解析には、は京都大学黒橋研究室の JUMAN をはじめ、奈良先端科学技術大学院大学松本研究室の茶筌 (松本、2000) や，google の工藤氏による

MeCab など、フリーのソフトウェアがある。前述の KHcoder，TinyTextMiner などでは、こういった形態素解析のツールを組み込んでいるので、ツールの存在を意識しなくても分析を進めることができる。

構文解析

形態素に分割された文章は集計の際には有効だが、文章の意味にまで踏み込んで分析を行う際には不十分である。こういった場合には、係り受け関係など構文について検討する必要がある。 構文解析に用いられるソフトウェアとしては、京都大学黒橋研究室の KNP(黒橋、2000) や、google の工藤氏による cabocha(工藤・松本、2001) が挙げられる。これらは、JUMAN や MeCab 同様フリーソフトウェアである。

分析の第一段階は，単語の頻度の集計である。集計方法は、大きく分けて二通りある。一つは、文章の中で単語が出現した個数を集計する方法である。この方法では、一つの文章である単語が複数回出現した場合、それぞれを出現回数としてカウントする。もう一つは、文章の中で単語が出現したか否かを集計する方法である。この方法では、一つの文章の中である単語が何回出現したとしても、一回としてカウントする。単語の集計により、分析対象となる文章の特徴を大まかに把握することが出来る。

共起

意味の分析を考える場合、文章や段落内での共起関係の分析も有用である。これは、単語同士の分割表を作成する形で集計を行う。前述の頻度集計が一次元の集計であるのに対して、分割表の作成は、二次元の頻度集計と考えることもできる。

統計解析

テキストデータの場合、単語を用いて集計を行なっても、出現単語数が多くなりがちであり、そのため、単純に結果を見るだけでは、有効な知見を得るのは難しい。そこで行われるのが統計的手法を用いた分析である。たとえば、書いた人の性別や年齢といったテキストの属性と出現単語を用いたコレスポンデンス分析、あるいは個々の文章と出現単語を用いた数量化三類、それらの結果を用いたクラスター分析が考えられる。それ以外にも、単語間の共起性を見る多次元尺度法やネットワーク分析、また、SVM のような機械学習による、テキスト分類などもよく行われている。

Japanese text b)

明治維新：近代国家への歩み [23]

黒船来航と不平等条約

19 世紀になると、産業革命に成功した列強諸国が市場を求めてアジアへ勢力を広げ、日本近海にも異国船が出没し、上陸して通商を求めてくるようになった。江戸幕府はその要求を拒んできたが、1853 年、米国東インド艦隊司令長官ペリーが強硬な態度で開国を迫った。幕府の首脳部は開国やむなしと判断、翌 1854 年、日米和親条約を結んで下田と箱館（函館）の港を開いた。その後、英国、ロシア、オランダとも同様の条約を結んだ。

さらに 1858 年には日米修好通商条約を結んだが、条約には一方的な最恵国待遇、領事裁判権、協定関税制度など不平等条項が盛り込まれた。領事裁判権は罪を犯した外国人を駐日領事が裁く制度で、日本人は外国人を断罪できない。さらに日本側に関税自主権はなく、税率は非常に低く設定された。結果、生糸や茶が大量に海外へ流出して国内で品不足となり、連動して諸物価が高騰。その一方で、安い綿織物が日本に流れ込んだため、綿作農家や綿織物業は経済的に大きな損害を被った。

幕府の失墜と日本の混乱

こうした開国による混乱は外国人に対する憎しみとなり、孝明天皇が大の異人嫌いだったこともあり、天皇の下に結集して外国人を追い払えという尊王攘夷運動が高まった。大老の井伊直弼は運動を徹底的に弾圧（安政の大獄）したが、1860 年、登城途中で攘夷派の浪士たちに暗殺（桜田門外の変）された。幕府の最高権力者が殺されたことで幕府の権威は失墜、尊攘派（主に長州藩士）が朝廷を牛耳るようになった。

しかしその後、尊攘派は会津・薩摩藩など天皇の権威を借りて幕府を立て直そうとした公武合体派によって朝廷から駆逐された（8 月 18 日の政変）。これに激した長州軍は、大挙して京都御所に乱入しようとしたが、会津・薩摩軍らに撃退され、朝敵として幕府の征討を受けるはめになった。

外国の脅威と秘密の軍事同盟

薩摩藩は薩英戦争、長州藩は英国・米国・フランス・オランダからなる四国艦隊下関砲撃事件を経験し、外国の強大さをあらためて実感。攘夷は不可能だと悟り、日本が外国の植民地にならないためには、素早く近代国家をつくる必要性があると痛感したのである。かくして 1866 年、薩長両藩は秘密の軍事同盟（薩長同盟）を結んだ。

同年の第 2 次長州征討では、薩摩藩は参加を拒否、密かに長州藩に大量の武器を送って支援した。結果、幕府の征討軍は長州一藩に敗北したのである。これを機に倒幕の勢いが一気に加速していった。

江戸幕府政権を返上するも、戦いが続く

対して将軍慶喜は1867年10月、平和的に政権を返上（大政奉還）し、新たに誕生する朝廷の新政権に参加しようともくろんだ。しかし薩長両藩はあくまで武力による倒幕計画を進め、同年12月、朝廷でクーデターを起こして強引に王政復古の大号令（新政府樹立宣言）を出させ、同夜の会議（小御所会議）で倒幕派は、慶喜の内大臣の職を奪い徳川家の領地を朝廷に返還させるべきだ（辞官納地）と主張。土佐・越前藩ら公議政体派（穏健派）の反対意見をねじ伏せた。

これにより薩長両藩は旧幕府方の暴発を狙ったわけだが、慶喜は二条城からおとなしく大坂城へ移り、事態を静観した。その後、新政府内で公議政体派が巻き返しを図り、倒幕派が失脚して慶喜の入閣が決定する。が、倒幕派の西郷隆盛らが浪人を送って江戸の治安を乱し、怒った旧幕府方が薩摩藩邸を焼き打ち。これを知った大坂城の旧幕臣は激昂し、彼らを抑えきれなくなった慶喜は京都への進撃を認めてしまう。こうして旧幕府軍と新政府軍（主に薩長両藩）の戦い（鳥羽・伏見の戦い）が起こり、旧幕府軍は敗北し慶喜は江戸へ逃亡した。

明治政府の始まり

新政府軍は5万の大軍で江戸を包囲したが、旧幕臣・勝海舟と新政府軍の西郷隆盛の会談により、江戸城を無条件で引き渡すことで総攻撃は中止され、慶喜の一命も保証された。一方、朝敵となった会津藩の救済を嘆願していた東北諸藩は、奥羽越列藩同盟を締結して新政府への敵対を明らかにし、各地で新政府軍と同盟軍の戦いが始まった。

この間、新政府は五箇条の御誓文を発して「開国和親、公議世論の尊重」方針を明らかにし、政体書を制定して米国の憲法を参考に三権分立制の新政府の政治組織を整えた。さらに天皇を江戸城に移して皇居とし、江戸を東京と改め、元号も明治に変えた。

旧幕府軍最後の抵抗

1869年5月、蝦夷地の箱館五稜郭を拠点とする榎本武揚ら旧幕府軍が新政府に降伏（戊辰戦争の終結）したことで、新政府は日本全土を統一した。そこで新政府は同年、諸藩の土地と領民の返還（版籍奉還）を大名に命じた。ただ、これはあくまで形式的なもので、藩主は知藩事と名を変え、そのまま領内の政治を執り続けた。

なお、戊辰戦争で戦ったのは諸大名の藩士だったので、その多くが国元へ戻り、新政府はほとんど軍事力を持たなかった。このため第二の戊辰戦争を想定して各藩はすさまじい軍事改革を始めた。紀州藩などは徴兵制度を創設、プロシア式の2万の近代的軍隊を作り上げた。

明治新政府による革命的政策「廃藩置県」

まさに革命的な政策だったことから、木戸や大久保らは大きな抵抗がある予想したが、意外にも騒動は起こらず、すんなり廃藩置県が達成された。その要因は、藩士の家禄（給与）や各藩の借金は新政府が請け負うと表明したことが大きかったようだ。いずれにしても藩は地上から消滅し、新政府は国内の政治的統一に成功したのである。以後、新政府は短期間にすさまじい社会変革を断行していった。できるだけ早く近代化し、富国強兵を達成して列強による植民地化を防ぐためだった。

士族の反乱から自由民権運動へ

江戸時代の税収は農民からの年貢が主で、収穫の豊凶によって毎年の歳入が大きく変化した。そこで新政府は、年貢負担者を土地所有者と認定して地価を記した地券を発行。1873 年、地価の 3 %を地租（租税）とし、土地所有者に納入を義務づけた。税率は収穫の豊凶にかかわらず一定とし、納入方法は金納とした。これにより近代税制が確立され、国家の歳入は安定した。また、国民はすべて平等とする四民平等政策を進め、20 歳以上の国民（男性のみ）に徴兵検査を課し、合格者の中から 3 年間の兵役を課した。こうして武士ではなく、国民からの徴兵による常備軍を創設したのだ。

これに不満を持ったのは士族（元武士）であった。廃藩置県で主家を失い、さらに代々の禄（給与）も金禄公債（一時金）と引き換えに 1876 年に打ち切られた（秩禄処分）。四民平等と廃刀令によって、苗字帯刀の特権も失った。このため同年から士族の乱が相次ぎ、翌 1877 年には西郷隆盛が挙兵（西南戦争）した。だが新政府軍は全力で反乱を平定、以後、武力による政府の打倒は不可能となった。代わって盛り上がったのが自由民権運動だった。

土佐出身の板垣退助らが政府の薩長藩閥独占体制を非難し、選挙で選ぶ民撰議院（衆議院）を開設して民間人を政治に参加させよと主張したのが始まりだ。運動の担い手は不平士族から豪農、さらに国民一般へと広まっていった。

天皇主権の大日本帝国憲法を制定

こうした中、政府は憲法の制定へと動き始める。日本が近代国家として国際的に認められ、不平等条約を改正してもらうためにも憲法の制定は急務だったが、政府を突き動かした最大の要因は自由民権運動の高まりにあった。民権家は政府に立憲政体の樹立と国会の開設に加え、憲法の制定を声高に叫び、自らも私擬憲法を盛んに作り始めた。その多くは国民の権利を重視した民主的な内容で、フランス流の急進的な案もあった。

政府の高官は、天皇制と藩閥体制を強化する憲法を模索していたが、やがて政府内部からも大隈重信のようにすぐに英国流の斬進的な憲法を作るべきだという声が上がった。動揺した高官たちは 1881 年の政変で大隈派を政府から追放、憲法研究のため伊藤博文をヨーロッパに派遣した。欧州各国の憲法を比較検討した結

果、伊藤は君主権の強大なドイツ流の憲法を参考にすることを決め、帰国後、日本の実態に合うように工夫と修正を加え、草案を枢密院に提出した。枢密院とは、憲法草案を審議するために置かれた天皇の最高諮問機関である。

枢密院では、天皇出席の下、何度も法案をめぐる議論が交わされ、1889 年 2 月 11 日、欽定（天皇が定めた）という形をとって、大日本帝国憲法が発布された。この憲法の特徴は、神聖不可侵の天皇が主権を持ち、「天皇大権」という絶大な権限を有するところにあった。天皇は統治権の総攬者であり、軍隊の統帥権を握り、かつ内閣の任免権を有するとされた。しかし、憲法の範囲内という制限付きながら、信教・職業・言論の自由など、国民の権利がかなり広く認められた。このような自由権を憲法に入れたのは伊藤の要望だったらしい。

藩閥政府の中心的人物でありながら、伊藤は後に立憲政友会（政党）を立ち上げて政党内閣を作ろうとしており、比較的リベラルな思想の持ち主だった。しかも伊藤は憲法の解釈にも幅を持たせたので、最大限に民主的な解釈をとれば、美濃部達吉のような天皇機関説（天皇は国家の最高機関とする説）に到達するし、言葉どおりに解釈すれば天皇至上主義に行きつく。そして実際、前者が大正デモクラシー時代を築き上げ、後者が軍国主義という暗い時代を生み落としたのである。

いずれにせよ、憲法の制定によって日本はアジアの中でいち早く近代国家の体裁を整えることができたのである。