Multiple Usage Corpus Generation Based on News Websites with URLs Extractions, News Selections Skills

Yingying Lao[1], Takumi Hosokawa[1], Tong Wu[2], Dongli Han[1]
han@chs.nihon-u.ac.jp *twu665@usc.edu*
[1]Nihon University *[2]University of Southern California*

Abstract

Corpora have been widely used in natural language processing research of various topics. In general, scholars would apply a certain corpus in each distinct study. Even though there have been a few corpora that could be applied to multiple fields and some of them are generated from the news, automatic news corpus generation system has not been proposed by previous scholars. There are previous studies which used a specific news corpus to resolve an individual problem, however, none of them could offer a corpus that is applicable to different situations. In this paper, we design a system to defeat this shortage, so that the system could automatically collect news from news websites and generate corpora. To achieve this goal, we extract news URLs from different websites and assigned forward, then select adequate news URLs and their contents. In order to apply our system to multiple news websites, we create a common module for extracting news titles and contexts from websites and use the contents to generate corpus. We test this model on 24 news websites to estimate the response. Experimental results have shown the effectiveness of our approach.

Keywords*: Corpus, News Websites, Natural Language Processing,*

1. Introduction

Corpus, a text resource, contains a mass of structured texts written by natural languages and could be used for statistical analysis and hypothesis testing [1]. Sometimes, it is difficult to use existing corpora with an appropriate text extraction system, people could generate a suitable corpus for their studies. Numerous corpora are utilized frequently in language research and information processing. The twitter corpus, which has been collected from Twitter, is used to analyze the usage of colloquial vocabularies [2]. Likewise, in linguistic research, scholars create dialect corpus to investigate sentence structures and patterns which contain dialect features [3]. By using corpus based on novels, people could study different uses of one word in diverse contexts [4]. Additionally, corpora are widely used in research of compound phrases with equivalent meanings [5]. All the information above shows the necessity of conducting the research. From previous works, we could conclude that considerable numbers of corpora perform relatively well in their fields. However, even if several studies have proved the feasibility of applying one corpus to various topics, no multi-subject corpus has been brought forward until now, which lead to the utilization of several corpora in the same experiment. This situation inevitably results in the increase of their cost and weakens its compatibility. Therefore, building a state-of-the-art news corpus with numerous categories of the news information might be a solution to current problem. In this paper, a universal corpus has been designed whose core task is to extract accurate news bodies from varied layers of one news website. In the following sections, a news extraction system with its experiments and corresponding tests will be introduced. The news extraction system would apply to multiple news websites.

In order to extract relevant websites and build up news corpus, two challenges should be confronted. One is news webpage URLs accurate extraction and noise elimination. The other challenge is the system's adaptation capability. Therefore, an algorithm called prescribed module is designed to collect valid news web pages and get rid of those noisy pages, such as advertisements, videos, and navigation pages. Other than that, due to the diversity of news websites' layouts and constructions, a more universal algorithm should be proposed to extract news headlines and contents. For validity reason, verification test would be applied to the algorithm with 24 news websites. In this paper, we first introduce the system flow in Section 2, then make an evaluation of our URLs extracting methods among 24 testing websites in Section 3, and examine two selection methods of news URLs in Section 4, then choose relevant and non-repetitive news context from web pages in Section 5, finally, we make a conclusion about this process which could generate a universal corpus based on news, meanwhile propose some comments and suggestions for further studies.

2. System Flow

Currently, there exist several website extraction tools, however, they could not satisfy the demand in our study. In this paper, the ultimate goal is improving the existing system to be more suitable for extracting news rather than pictures. One of the well-established systems is called Webstemmer. Webstemmer is a machine learning tool which could automatically distinguish news pages layout and extract news contexts from target websites [6]. Although Webstemmer performs well in previous works, it does not maintain the behavior if the page layout has been modified and shown differences from the original layout. This phenomenon implies one of the features of news websites. News websites contain diverse information, such as site introduction, news, and advertisements. With this feature, it is inevitable to extract some irrelevant pages when we are using the existing tool.

To amend the shortages of Webstemmer, we propose a news extraction system which consists of 4 procedures. First of all, we import the news website's homepage URL and the user's expected number of searching layers into the system. Here, the searching layer is the webpage that could not only be accessed through hyperlinks on the homepage but also lead to other web pages by its hyperlinks. Next, according to the expected number, the system detects and collects a large number of URLs from the given news website. Then, with different models, the system reflects its estimations and selections of the news pages URLs exclusively. Finally, from those selected URLs, the system would extract their headlines and contexts. With those extracted text resources, we could generate a news corpus from the target news website.

3. The extraction method of URLs

The first procedure of our system is to collect URLs from the target news website. We observe that news website's homepage holds a large number of pages, and each page contains various hyperlinks. Only some of those hyperlinks are news pages that we expect to extract. Moreover, news pages could link to other correlate news pages. Due to the multiple correlations among page layers, we could acquire abundant URLs from the target news website efficiently. Thus, we introduce a variable N-layer to represent the number of searching layer for the system. It helps users to find and collect certain layers of URLs from one target news website. From assigned target homepage and N-layer, the system seeks and gathers URLs automatically and avoids gathering duplicate URLs.

To illustrate this system, we now provide an example where N-layer equals three. The system starts to collect URLs from the first layer which is the homepage of the target website given by the user. Then, we determine each collected URL from the last step as the second layer target websites and repeat the algorithm to collect new URLs. For simplicity and efficiency reasons, the system would abandon the URL if it is not unique.

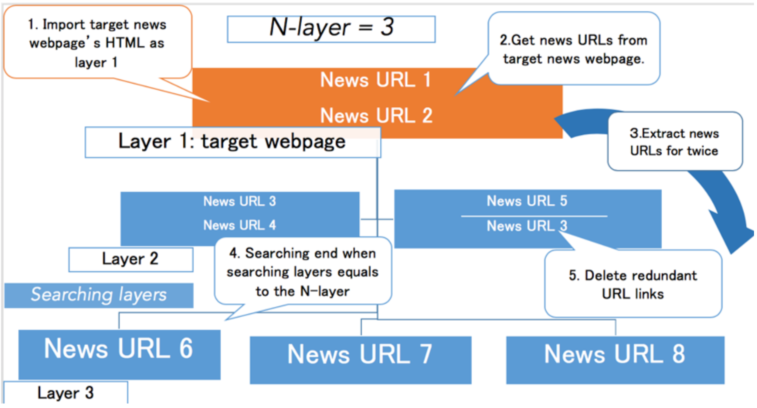
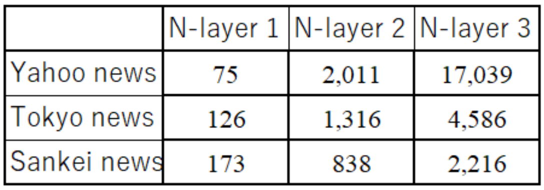

1. Import target news webpage's HTML as layer 1
N-layer = 3
2.Get news URLs from target news webpage.
News URL 1
News URL 2
Layer 1: target webpage
3.Extract news URLs for twice
News URL 3
News URL 4
News URL 5
News URL 3
Layer 2
4. Searching end when searching layers equals to the N-layer
5. Delete redundant URL links
Searching layers
News URL 6
News URL 7
News URL 8
Layer 3

Figure.1 URL extraction system with N-layer equals three

After constructing the architecture of the experiment, we choose 3 news websites (*Yahoo news, Tokyo news, and Sankei news*) to examine our proposal. While the N-layer is assigned as 3, the number of collected URLs are shown in table.1. It is important to denote that every statistic in table.1 is representing the number of additional URLs from each layer. For instance, the numbers of N-layer 2 show how many URLs are extracted from layer 2 but not extracted from layer 1. Similarly, the numbers of N-layer 3 are displaying the numbers of unique URLs extracted from N-layer 3. When the N-layer increases, the total number of collected URLs for each news website grows up as well. Taking the Yahoo news website as an example, there are 75 URLs gathered in N-layer 1, 2,011 URLs collected in N-layer 2, and 17,039 URLs caught in N-layer 3. Each N-layer (except N-layer 1) could collect a bigger number of URLs than previous N-layers since news websites would like to list news articles from an earlier time in the recommendation section. By introducing N-layer, we could gather a larger number of URLs with less cost.

Table.1 the Number of collected URLs from 3 news websites



	N-layer 1	N-layer 2	N-layer 3
Yahoo news	75	2,011	17,039
Tokyo news	126	1,316	4,586
Sankei news	173	838	2,216

This phenomenon appears in all of the three tested websites, indicating that this phenomenon might be universal. Although we try to use regression methods for the data of news websites to identify their regression functions for further studies, it is difficult for us due to the insufficiency of data. However, we would suggest that none of those functions is a linear model and they might be some kind of exponential function since we observe explosive increases from N-layer 3.

By analyzing the content of extracted URLs, we recognize that some of the extracted URLs are news from an earlier time when we proofread the content of them. However, there are still a massive number of irrelevant websites which add extra noise in our dataset. Therefore, it is necessary for us to come up with methods to select news URLs.

4. Selection methods of news URLs

To select news URLs from the total collection, we propose two algorithms called the normal representation algorithm and the statistical information algorithm.

4.1 Normal representation algorithm

From URLs collected above, we identify that the news URL is written with structured patterns and contents, such as serial numbers or dates. Therefore, we design the following 3 criteria for examining URLs. If an URL satisfies at least one of the following criteria, the system would extract and collect it as a new URL.

a) **The URL includes numbers which represent publication dates;**
b) **The URL contains hexadecimal character strings;**
c) **The URL's character string consists of alphabets and numbers.**

We arrange a test to verify the reliability of the normal representation algorithm. First, we choose 10 different news websites and gather all URLs from their homepages. Then, we list all of the news URLs from those homepages manually and count the number of news URLs as N-news. After determining the standard answer (N-news) for this experiment, we initiate the process of gathering experimental objects which are qualified URLs according to our normal representation algorithm selection.

All of those URLs should be gathered by the normal representation algorithm on the same day for the sake of fairness. We request the system to extract news URLs from each target website with assigned parameters. To evaluate the effectiveness of our algorithm, we use precision and recall to get quantitative measurements. Precision is the ratio of extracted news URLs number by normal representation system after manually checking and the total number of extracted URLs by the algorithm. It illustrates the accuracy of our algorithm by its value. When the precision gets closer to 1, it represents that the algorithm has a higher accuracy. Recall means the ratio of the news URLs gathered over N-news. In statistics, recall represents the ability to identify true positives and it would approach 1 when the algorithm identifies true positives progressively. When the recall gets bigger, the precision decreases due to the efficiency. Our results reflect this pattern in 10 testing websites.

The results of the test websites are shown in Figure.1. We observe the precision and recall of most of the websites exceed 90 percent, except the Nikkei website whose precision merely attains 67 percent. We have several assumptions for Nikkei's result listed as following. Since there are a large number of news pages on Nikkei, the correct data of news are enormous which adds challenges to the test. Then, some of the news URLs are formatted with character strings which do not adjust to our criterion of normal representation. For those reasons, the system picks out wrong URLs and devotes in low precision. Nevertheless, we argue that the normal representation algorithm could be applied to most of the news websites due to the high average precision and recall which both come up to 95 percent.

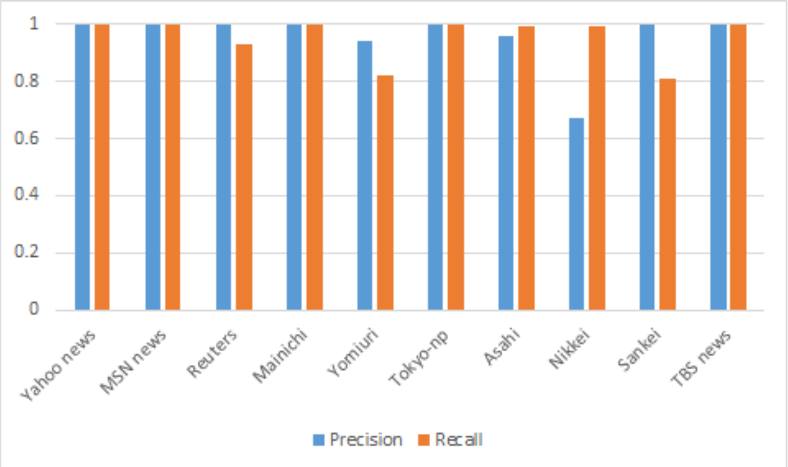


Figure.1 Test results of 10 news websites-normal representation algorithm

4.2 Statistical information algorithm

We are aware of another common phenomenon about URLs of news pages besides the normal pattern mentioned above. Statistically, it's feasible to find out that news pages use longer URLs and other pages. Therefore, we use the length of the URL as a clue to design the statistical information algorithm.

The algorithm is composed of two steps. First, according to the length of the URL, we classify the gathered URLs which are extracted from chapter 3 into certain groups. With different classification methods, we break the data into different groups with different widths, so that we could separate news apart from others. Through analyzing the classification, we catch length range of news pages' URLs. Second, the system distinguishes news pages' URLs from others within the length range.

We make Figure.2 to illustrate the distribution of length classification of gathered URLs. The X-axis represents different classes of the URLs' length starts from 0 with the break equals to five. And the Y-axis represents each the number of news pages' URLs. The blue bar represents URLs of news pages and orange bars represent URLs of other pages.

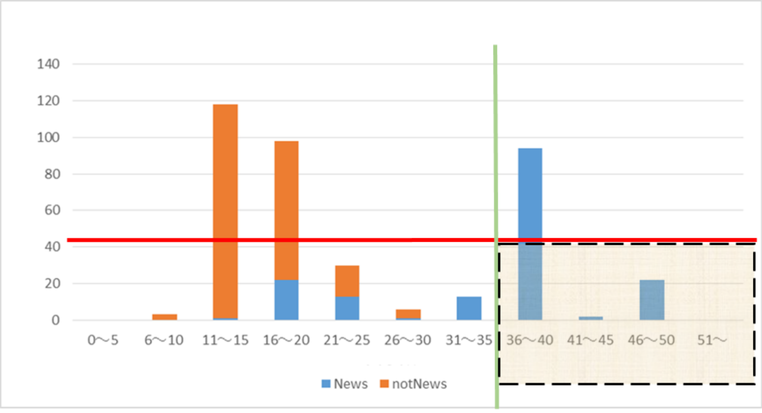
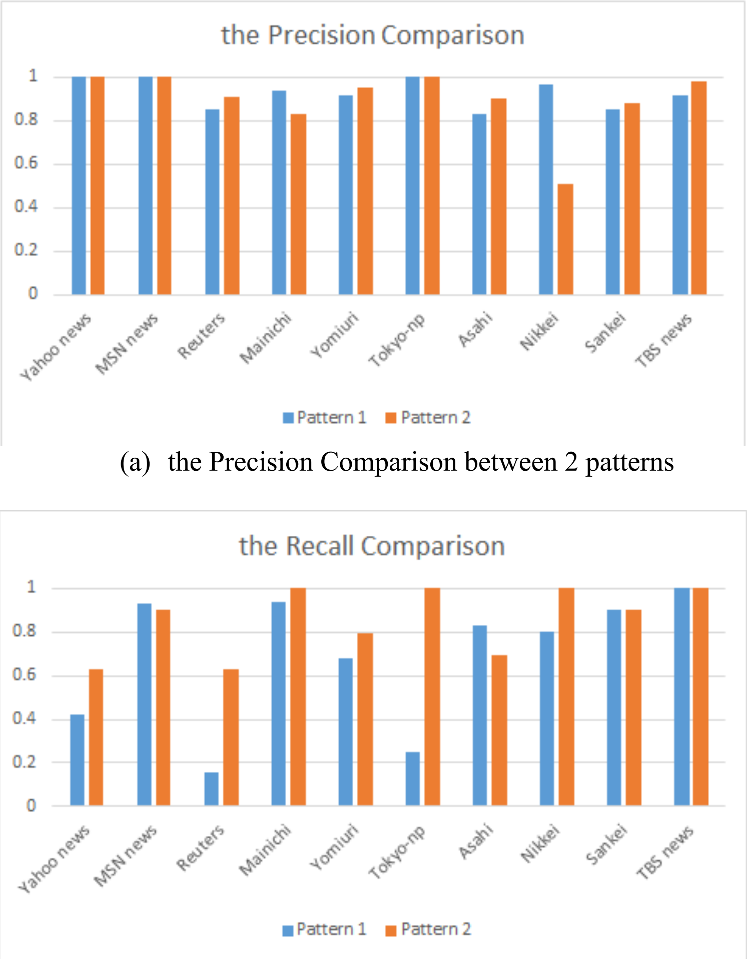


Figure.2 The distribution of length of news pages’ URLs

After analyzing the distribution of long URLs, it's not complicated to find those URLs contain certain directories or serial numbers. For every URL, there are several segments which represent different meanings and they are always separated by slash lines. We observe two structural patterns which could be helpful for determining the website types. One is the length of the involved directory, and the other one is the length of the longest segment in a certain URL. For examining the efficiency of two proposals, we have extracted URLs of news pages from 10 different news websites' homepages which have been used above in section 4.1. Here, we also compute the precision and recall of the statistical information algorithm. Furthermore, we compare two precisions from different algorithms to evaluate those two algorithms. From this process, we could analyze the advantages and shortages of both processes and combine them together to form an improved algorithm.

Figure.3 shows the testing results of the statistical information algorithm. Pattern 1 represents the length of the involved directory and pattern 2 represents the length of the longest segment. In the normal representation algorithm, the precision would be probably lower than average for some specific websites. However, from figure.3. (a), we observe that all of the test websites' precisions are over 80 percent when we are using statistical information algorithm. Other than that, from figure.3. (b), we could find the average recall of pattern 1 is 69 percent and pattern 2 is 85 percent. In statistical information algorithm, both of patterns' average recalls drop off extremely comparing with the normal representation algorithm. We conclude the reason to outliers because there are some news pages have a shorter length of URLs than the lower bound of our range. For this reason, the system fails to extract some news pages' URLs and brings the low average recall.



(a) the Precision Comparison between 2 patterns

(b) the Recall Comparison between 2 patterns

Figure.3 Test results of 10 news websites - statistical information algorithm with 2 patterns

Pattern 1 could amend the existed shortage in the normal representation algorithm whose extracted news URLs precision is lower than average. However, the average recall of pattern 2 performs better than pattern 1. In Figure.3. (a), when URLs of news pages are identified by their quantity, the precision of the statistical information algorithm would go up. On the contrary, the recall would be improved when a little number of news pages' URLs are attained, which is shown in Figure.3. (b). This phenomenon is observed from Yahoo news. Consequently, we design a system that runs two patterns of statistical information algorithm to gather news pages' URLs. Based on the number of news pages' URLs gathered from each pattern, the system would return the result of extracted URLs that are satisfied user's preference with a higher precision or a better recall.

4.3 Summary

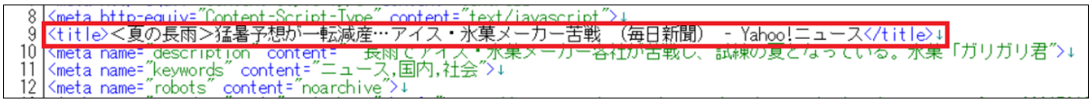
For extracting URLs of news pages, two algorithms are integrated into our proposed system for user's selection. For common news websites, the normal representation algorithm could gather news pages URLs with a high precision and a compatible recall. However, the recall would be weakened for some peculiar news websites. In the statistical information algorithm, the precision and recall are quite well for those collected news pages URLs from every news website. Based on the user's choice of pattern, the result of obtained news pages URLs complies with excellent precision or exquisite recall.

5. Extract news text

In this part, we would illustrate how does the proposed system extract information which contains the headline and news' paragraphs from each selected news pages URL. In this section, we design an algorithm and get prepared for corpus generation with two steps. First, due to the special structure of HTML web pages, the system would grab the headline and the main context of news by searching those specific tagging patterns. Second, since the captured context contains abundant irrelevant information, so it's necessary to remove them for both efficiency and correctness. Final, the product from those two steps would be used for generating a corpus of news.

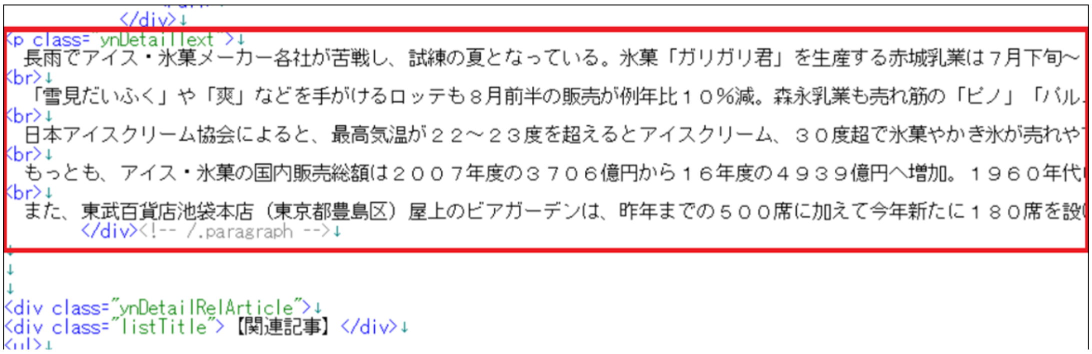
5.1 Certain pattern extraction

From URLs obtained above, we are able to access page sources of each URL. In the page source, we could read all of the HTML tags of one web page. Then, we design a system to distinguish the news body based on HTML tags and name the system as Certain Pattern Extraction. In general, web pages are made up with different HTML tags in fixed forms, such as headlines would be enveloped by two tags. The tag <title> would be placed in front of the title and the tag </title> would be inserted after the title. Other than titles, both texts and pictures have their own tags such as "<p></p>" and "<pic></pic>" which would be placed at the same position as title tags. Although HTML tags are adequate for distinguishing specific substance of web pages, we do need to add some extra restrictions when we extract the text of news. Restrictions are defined as particular expressions. For example, since stop words or quotation marks are always applicable to the end of texts, they could help us to locate news texts. We could select texts which are not only close but also in front of those features. Comma or hiragana usually appears in the news body since Japanese news uses them more often comparing to advertisements. So, they could also be our target to locate news bodies. There are samples of extracted pages shown in Figure.4 and Figure.5.



```
<meta http-equiv="Content-Script-Type" content="text/javascript">↓
<title><夏の長雨>猛暑予想が一転減産…アイス・氷菓メーカー苦戦 (毎日新聞) - Yahoo!ニュース</title>↓
<meta name="description" content="長雨でアイス・氷菓メーカー各社が苦戦し、試練の夏となっている。氷菓「ガリガリ君">↓
<meta name="keywords" content="ニュース,国内,社会">↓
<meta name="robots" content="noarchive">↓
```

Figure 4. A sample of HTML tag- title



```
        </div>↓
<p class="ynDetailText">↓
 長雨でアイス・氷菓メーカー各社が苦戦し、試練の夏となっている。氷菓「ガリガリ君」を生産する赤城乳業は７月下旬～
<br>↓
 「雪見だいふく」や「爽」などを手がけるロッテも８月前半の販売が例年比１０％減。森永乳業も売れ筋の「ピノ」「パル
<br>↓
 日本アイスクリーム協会によると、最高気温が２２～２３度を超えるとアイスクリーム、３０度超で氷菓やかき氷が売れや
<br>↓
 もっとも、アイス・氷菓の国内販売総額は２００７年度の３７０６億円から１６年度の４９３９億円へ増加。１９６０年代
<br>↓
 また、東武百貨店池袋本店（東京都豊島区）屋上のビアガーデンは、昨年までの５００席に加えて今年新たに１８０席を設
      </div><!-- /.paragraph -->↓
↓
↓
↓
<div class="ynDetailRelArticle">↓
<div class="listTitle">【関連記事】</div>↓
<ul>↓
```

Figure 5. Sample of HTML tag- text

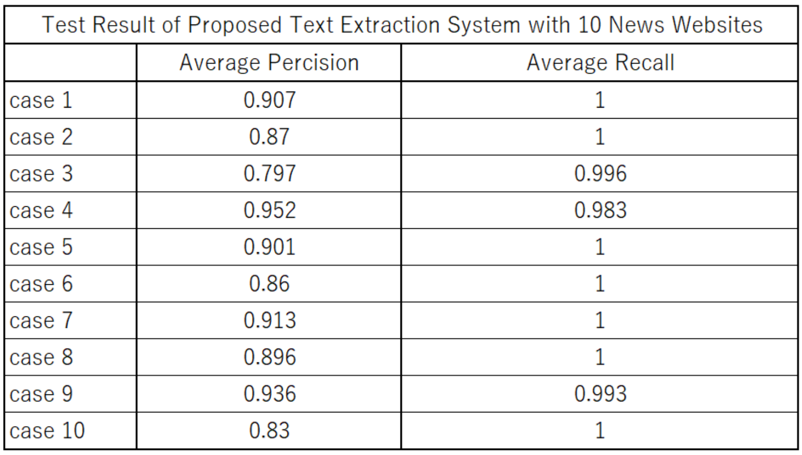
5.2 Remove irrelevant content

Even though we have been applied the Certain Pattern Extraction to select the main body of news, there are still remaining plenty of irrelevant contents such as advertisements, copyright links and so on. We are not satisfied with the result when it still contains abundant irrelevant contents. In order to manage diverse news websites, the Certain Pattern Extraction conserves some room for different structures and features. To eliminate irrelevant strings, we try to spot some common traits for the same website. We are conscious of irrelevant contents recurring frequently in the same website. For instance, the copyright string or warning text of website appears around the news content regularly. Therefore, our system can target and delete meaningless recurrence texts among gathered news.

5.3 Evaluation

To prove the efficiency of our proposal, we choose 10 news websites and conduct a test. For each website, the system randomly extracts 10 pieces of news' content and its title with mentioned procedures. Then, we manually draw-out the contents from those 10 news websites as the standard answer to compare with the result from our designed system. Finally, we compare the system extraction with the news content obtained by the human. According to the results of the comparison, we could calculate the precision and recall of the proposal, and verify the effect of out irrelevant content removal procedure. The precise represents the accuracy which is the ratio of valid content extracted from the system and our standard answer. The recall shows the ratio of success which is represented by the portion of content extracted entirely from each news' URLs. In general, we consider the matching rate on sentence level as a requirement. Therefore, we calculate the precision and recall without news headline in order to keep the requirement mentioned above. Additionally, when the whole content is extracted, we could do the calculation including the headline. Table.4 illustrates the average result of precision and recall for 10 websites in 10 different experiments. Although we observe the average precision and recall are both exceed 90 percent, the precision is below 50 percent for some news websites and the recall could not achieve 100 percent.

Table.4 the Result of Proposed Text Extraction System with 10 News Websites



Test Result of Proposed Text Extraction System with 10 News Websites

	Average Percision	Average Recall
case 1	0.907	1
case 2	0.87	1
case 3	0.797	0.996
case 4	0.952	0.983
case 5	0.901	1
case 6	0.86	1
case 7	0.913	1
case 8	0.896	1
case 9	0.936	0.993
case 10	0.83	1

We summarize the reasons for this phenomenon as follow. There are four aspects that interfere with the performance of precision. First, advertisements are treated as the content of the news website instead of being removed completely. Second, the same news gets extracted repeatedly. For example, the abstract of news is always composed of quotations from the main body of news which is the redundant part when we are extracting the news' body. Third, comments of pictures are picked out superfluously. Fourth, HTML tags are remained and collected as part of news content. Thus, we try to find solutions for solving those problems and increase the performance.

In our proposal, we assert excising advertisements through seeking recurrent content in massive news pages. However, the advertisement has not been removed successfully, since the difference of appearing frequency among various type of advertisements. Some of them only show up once in our testing websites which is hard to target and aim for the removal process. If we increase the quantity of viewed news pages, the problem could be solved. By adding more news for evaluation, low-frequency advertisements would be detected by the system and removed easily.

Next, we would like to focus on the redundant content. The duplicate texts extracted by the system are mostly derived from title and abstract of news. To deal with them, we consider to compare the duplicate content with the main body of news and delete the part which contains fewer words. For instance, when we compare each sentence with its counterparts in the news, we should use our system to identify those sentences which have high similarity with the target sentence. Then, we keep the sentence which has the longest length among all of the similar sentences. By this method, we could retain the most information with the least capacity.

After dealing with the duplicate texts, we start to manage the picture comments. Though pictures' comments have occurred constantly in news content with rigid forms, we intend to eliminate those comments by using HTML tags. As we mentioned before, pictures obtain their unique tags on the HTML page. We believe the comment would probably be close to the related picture. According to the HTML tag of pictures, the system could find out pictures' position in news content and identify the comment of a picture. Then, we could remove those comments efficiently.

Although HTML tags is an effective way to identify different contents parts, they might cause extraction mistakes. Such as content which is written by line separation tags(<p> and </p>) could make the system fail to identify the correct data. Therefore, we propose to modify the content in one line, where the content contains line breakers. The new contents would be presented as <tag>content</tag>. It is conducive to improve the system's collection ability of news and avoiding to extract HTML tags.

6. Conclusion

In this paper, we have proposed an approach to distinguish and extract news website URLs from target websites with a higher accuracy. Our approach mainly includes three parts: URLs extraction, news URLs selection, and news text extraction.

Comparing with previous studies where could only be applied to one specific website, our proposed system could be used for 24 different news websites, and extract related historical news websites through the given homepage URL and the layer parameter. Another difference between previous works and our system is that we design two different algorithms for news website URLs selection. Unlike other programs which just use one algorithm, this system combines two algorithms and merges their advantages together, meanwhile provides the autonomy for users to choose the fittest algorithm from their own perspectives.

In the URL extraction model, we innovatively introduce a theory such that derive news websites' URLs from the target website with a parameter provided by the user. With this theory, users are able to extract related news' URLs from hyperlinks on the current website. With different parameters, various related news and historical information are accessible to users. With this method, we collect explosive number URLs and data as long as the parameter gets bigger.

News page selection section from our suggested system is different from previous works. We realize that there are certain patterns in news' URLs, such that news URLs generally have longer lengths than other URLs and they also contain dates or times in their strings. First, we discover the pattern of length among several news URLs and try to pick out news websites by comparing their lengths with the standard news URL's length. However, some of the news URLs are outliers which are not satisfied with this criteria. It can lead to a lousy result during searching, therefore, we reveal another way to select suitable URLs. The second selecting method is to identify dates and times in URLs' character strings. We realize that neither method performs well when we analyze the precision and recall. Therefore, we combine them together to draw the selection range for news URLs which find a balance between pursuing high precious and recall.

Based on the tag feature from the source page, our news content selection method contains two steps. The first step is to locate and extract news body and headline, instead of using keywords, we use tags from the source page and pinpoint the content. In the second step after gathering all of the sentences, we delete irrelative materials and repetitive parts. We suggest four approaches to access the most simplified version of news body for corpus generation. Results of testing experiments reveal the effectiveness of our proposal.

There are still several ways to improve our approach to obtaining better performance. During our process, some certain websites are possessing relatively low accuracy no matter what kind of modification we apply to the system. In further studies, we might try to remove irrelevant content, such as comments and advertisements, before extraction with the tagging method. Another development we could make is applying deep learning for content classification. Deep learning could classify the extracted contents into different categories so that the users could choose to include or exclude specific categories.

7. Reference

[1] https://en.wikipedia.org/wiki/Text_corpus

[2] Ayaha_Osaki, Shohei_Karaguchi, Takuya_Osho, Toshiya_Sasaki, Yoshiragi_Kitagawa, Yuya_Sakizawa, Mamoru_Komachi, Building a Corpus for Twitter Japanese morphological Analysis, The Association for Natural Language Processing, pp.1-6, 2016.

[3] Soichiro_Hirota, Ryohei_Sasano, Oya_Takamura, Okumura, Construction of Dialect Corpus Collection System, The 27th Annual Conference of the Japanese Society for Artificial Intelligence, 2B1-3. 27 (p1-p4), 2013.

[4] Yanshan_Zhao, Semantic Usage Analysis of Modern Japanese Language, Language & civilization, no.15, pp. 61-p79, 2013.

[5] Tomoko_Ikeya, Difference Between the Japanese Compound Verbs "Das" and "Hajime": A Corpus-based Study, Theoretical and Applied Linguistics at Kobe Shoin: Talks, pp. 35-39, 2017.

[6] Webstemmer, http://www.unixuser.org/~euske/python/webstemmer/index-j.html#sample