Pavlo Bazilinskyy, Yuichiro Arai, and Solomon Adebayo

MSc Computer Science/Mathematics students at the University of St Andrews

Natural Language Processing and Text Mining

MINING THE OPINIONS OF JAPANESE TOWARDS CHINA AND SOUTH KOREA

Introduction

The relationships between Japan and China, Japan and South Korea have been dwindling over the last few decades. Two recent incidents further deteriorated the relationship over the last year. An announcement made by the governor of Tokyo that Senkaku islands would be acquired by the government from a private Japanese owner in an attempt to protect these islands from invasion by the Chinese activists triggered protests in numerous Chinese cities, leading to violent activities such as destruction of Japanese firms and cars in August 2012. South Korean president, Lee Myung-bak, declared that Japanese emperor Akihito must apologize from the bottom of his heart. Japanese considered it as a serious insult, and Japanese diet adopted a resolution against his speech. The aim of this study is to mine, analyze and process the opinions of Japanese towards China and South Korea with the help of manual opinion mining techniques.

Figure 1 shows the result of an opinion survey, which was conducted by the Genron NPO and China Daily. The attitude of the Japanese towards China and vice-versa were studied using the placement method during a period of time between April 2012 and May 2012, collection 2627 observations or opinions (N=2627). Results received from the surveys described above illustrate that 84% of Japanese have a negative impression about China and 64.5% of Chinese have bad impression about Japan [1]. Japanese opinion about South Korea and vice-versa were studied using the placement method from March to April in 2013 (N=2004). It was concluded that approximately 40% of Japanese have negative impression about China, and about 80% of South Koreans have bad impression about Japan. [2]

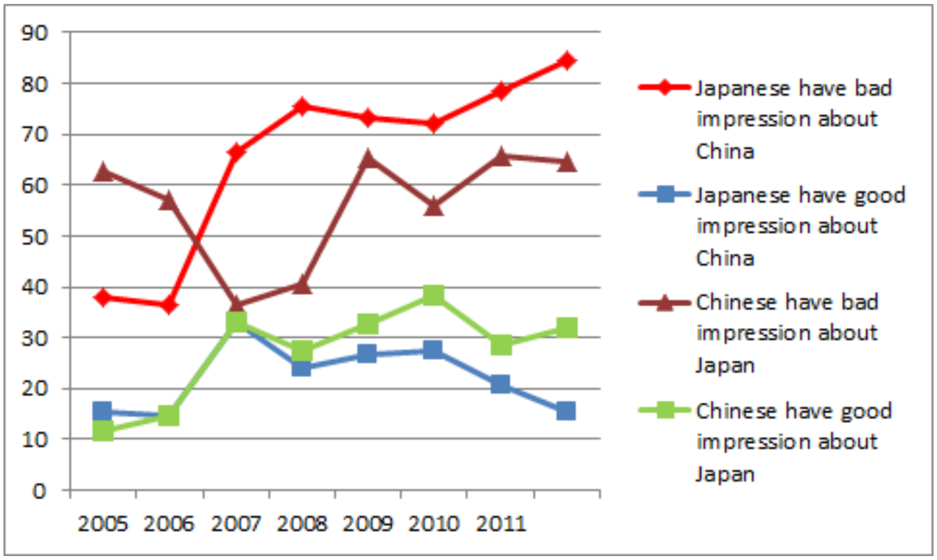


Figure 1. Relative Opinions about Japan, China and South Korea by their respective citizens [1].

Text Mining (TM) and Natural Language Processing (NLP) Algorithms

Text data mining and Natural Language Processing (NLP) techniques were applied in analysing results obtained from the Google search. NLP is an overlap between text mining [6], information retrieval [3], and text categorization [4]. Processing was done by Information Extraction (IE) algorithm of the Natural Language Processing algorithms concerning part-of-speech and sentiment analysis. We now define the terminologies used in these algorithms.

Terms are the words contained in sentences and are grouped into feature sets. The set excludes all stopwords. **Part of Speech (POS)** could fall into any of categories, namely; Nouns verbs, adjectives, adverbs, pronouns, prepositions, or conjunctions. The sentences extracted from the google documents contain numerous parts of speech, hence necessitating the application of POS algorithm implementations in our NLP toolkit [1]. **Part of Speech (POS) tag** is used to represent a feature. A popular approach is the Maximum Entropy POS tagger and the chunker, which is provided by Curran et al. in

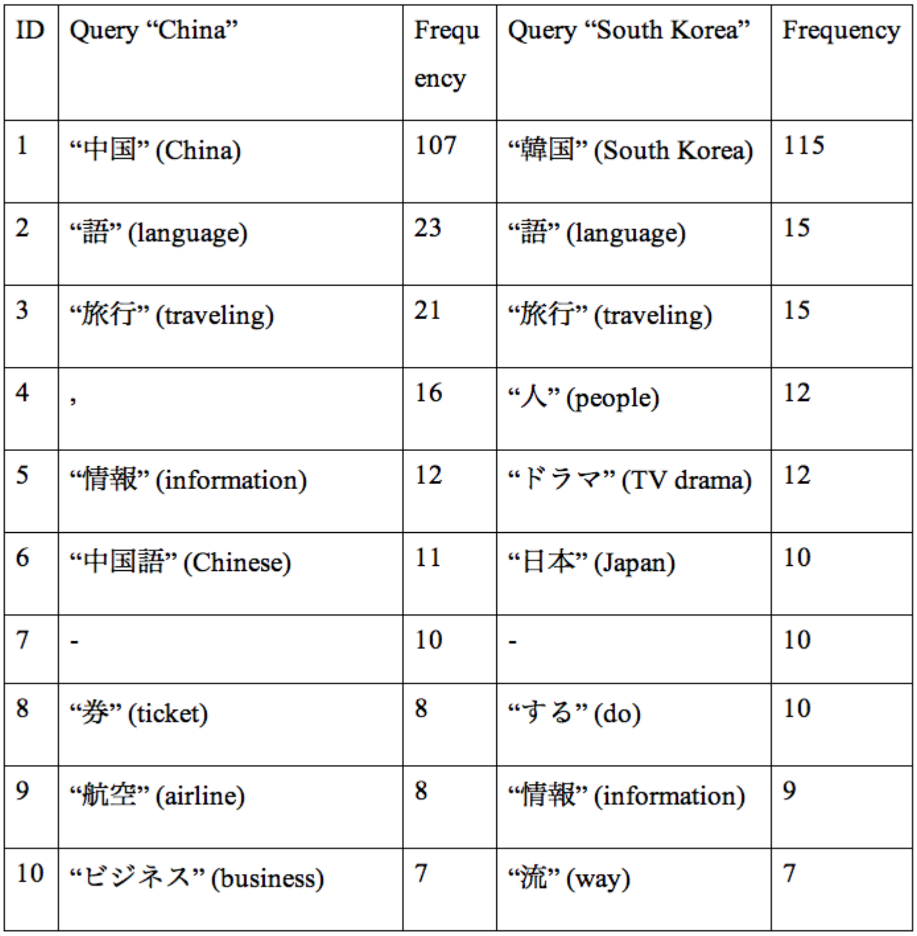
[5]. **The POS tagger** applies the aforementioned standard grammatical POS categories. POS tags include: *NP (Noun Phrase), VP (Verb Phrase), PP (Prepositional Phrase), ADJP (Adjective Phrase), ADVP (Adverb Phrase)*, etc.

Methods

In order to investigate the difference between Japanese impression of the Chinese and the Koreans on the Internet, a series of web search queries were performed at *google.co.jp* search engine. Sentimental analysis - that is, opinion mining - was performed on the obtained texts, and it was found that these web queries produced results of high sentimental values, and expressions containing strong impressions and opinions of Japanese towards the Chinese. Top 100 results received from the search in June 2012 were analysed and processed with the **mecab toolkit**. The names of data sources - the websites containing the articles of interests - were removed from the text data used for this study. It was expected that many sentences that are related to Japanese national interests (e.g. disputed islands, gas fields and copyrights) would be seen when a query "□□" (China) is entered in the Google's search engine. In contrast, it was expected that results obtained by searching for "□□" (South Korea) would be of the more negative sort due to a somewhat nationalistic view of Korea that predominates Japanese websites. The results obtained are in sharp contrast to those anticipated for these types of web searches.

Results

The results of the aforementioned experiment are presented in Figure 2. This table contains a list of ten most frequently used words across the search outcomes obtained from Google in response to queries "□□" (China) and "□□" (South Korea). The fourth and seventh most frequent word for "China" and the seventh for "South Korea" are all punctuation marks, which may be omitted.



ID	Query “China”	Frequency	Query “South Korea”	Frequency
1	“中国” (China)	107	“韓国” (South Korea)	115
2	“語” (language)	23	“語” (language)	15
3	“旅行” (traveling)	21	“旅行” (traveling)	15
4	,	16	“人” (people)	12
5	“情報” (information)	12	“ドラマ” (TV drama)	12
6	“中国語” (Chinese)	11	“日本” (Japan)	10
7	-	10	-	10
8	“券” (ticket)	8	“する” (do)	10
9	“航空” (airline)	8	“情報” (information)	9
10	“ビジネス” (business)	7	“流” (way)	7

Figure 2. Results obtained from running search queries.

These results show that the claim that the impression of the Japanese people on Internet towards China and South Korea is negative is incorrect. The top result, unsurprisingly, in both cases is the name of the country being searched. The second most frequent word indicates that the Japanese people are interested in learning Chinese and Korean languages. However, we would notice that when the sixth most frequent word for “China” is combined with the second frequent word, we obtain further support for the fact that Japanese are interested in learning Chinese [8]. Furthermore, the Figure shows a strong evidence that Japanese people are interested in traveling to China, i.e. the third, eighth and ninth frequent word obtained for “China” are all words that may be linked with traveling abroad. Results obtained for both countries contain a word “information”, which could be seen as evidence that Japanese often seek information about affairs that involve China and/or South Korea. Also, inclusion of “Japan” in results obtained for searching for “Korea” points out to the fact that interaction between these two countries is higher, when compared to involvement of Japan in affairs of China and vice versa. Inclusion of “TV drama” in what is retrieved after searching for “Korea” indicates that the film industry of South Korea has influence on Japanese market. Words like “do” and “way” may be omitted due to their broad use and inability to determine in their contexts.

Outlined results demonstrate that the claim that Internet source in Japanese language show that Japanese people have negative attitude towards Chinese and South Koreans is not valid. Words such as “war”, “violence”, “domination” etc. are not the most commonly used in association with “China” and “South Korea” search terms on the Internet in Japan.

Conclusion and Future work

In our study, we examined the impressions of the people of Japan towards China and the South Korea. The data sources for our experiment were Google search engines. Unstructured documents were obtained from performing certain query searches in Google. Natural Language Processing and Text Mining techniques were applied in performing opinion mining, extracting meanings and understanding from these documents. We performed sentence-level analysis of the documents in order to identify the Japanese opinions about the Chinese and South Koreans. Our results revealed that search outcomes currently available on the Japanese part of Internet do not contain any evidence of the negative attitude of people from Japan toward China and South Korea. Instead, search queries yield expected outcomes that focus on traveling, conducting business activities overseas, learning languages, etc.

In future, we intend to investigate the reverse case in which the opinions of the Chinese towards the Japanese are mined and analysed. We also would extend the study to include the South Koreans as subjects. We also wish to implement our analysis algorithms using Python programming language, which offers better performance, wider acceptance and natural language processing support, using tool such as the NLTK. In future work, our analysis would take its root at the article level rather than the web titles analyzed in this study. We also intend to adopt sophisticated sentiment analysis techniques in future work.

Bibliography

1. The eighth Japan-China joint opinion survey, - 8. [Online] Available at: http://www.genron-npo.net/pdf/forum2012.pdf, 2012.
2. Genron-NPO. The first Japan-Korea joint opinion survey, - 2-5. [Online] Available at: http://www.genron-npo.net/pdf/forum_1305.pdf , 2012.
3. Manning, C., Raghavan, P., & Schütze, H. Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2009.
4. Sebastian, F. Machine learning in automated text categorization. ACM Computing Surveys, *34*(1), -1–47, 2002.
5. Curran, J. R., & Informatica, D. Linguistically Motivated Large-Scale NLP with C & C and Boxer, (June), - 33–36, 2007.
6. Feldman, R., & Sanger, J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.
7. The statistics of international exchanges in high schools - 37,[Online] Available at:http://www.mext.go.jp/b_menu/houdou/25/04/__icsFiles/afieldfile/2013/04/03/1332931_01.pdf, 2011.