ID: 71

Proceedings of the International Workshop
on Modern Science and Technology
Kitami, Japan/September 2010

Developing a Large-Scale Corpus for Natural Language Processing and Emotion Processing Research in Japanese

Jacek Maciejewski*, Michal Ptaszynski†, Pawel Dybala†

*Independent Researcher
jacek.maciejewski@gmail.com

†Hokkaido University, Graduate School of Information Science and Technology
{ptaszynski, paweldybala}@media.eng.hokudai.ac.jp

Abstract There is a lack of web-based corpora for Japanese language. Existing ones contain insufficient amount of data, consist of improper word segmentation, have no clear structure and are unsuitable for the needs of emotion processing research. In order to overcome the abovementioned problems we built a new corpus. It is based on blog entries from Ameba blogging service (http://ameblo.jp/). The corpus contains sentences from common language with a variety of expressions and writing styles, which is important for a number of tasks in NLP. The original structure (blog post and comments) is preserved, thanks to which semantic relations between posts and comments are maintained. The corpus is intended to serve as a base for compiling more detailed subcorpora. Its usability in this field was proven by using it to create emotion object corpus. The goal of this paper is to make reader aware of new linguistic resource that can be used in various NLP research projects.

1. Introduction

Some of the most vital linguistic resources in the field of natural language processing (NLP) are text corpora. This could include newspaper corpora, like Mainichi Shinbun CdRom [1], conversation corpora, like ASJ Continuous Speech for Research [2], or corpora of literature (Aozora Bunko [3]). The importance of corpora is widely recognized and numerous corpora have been compiled so far for many languages including Japanese. However, comparing to other major world languages there are little large corpora available to the public [4]. Moreover, almost all of them are solely based on newspapers, or legal documents [2]. Unfortunately they are usually unsuitable for the research on emotion processing as emotions are rarely expressed in this kind of texts. There are of course speech corpora, but due to the difficulties with compilation they are relatively small.

In research such as the one by Abbasi et al. [5] it was proved that public Internet services consisting of social networks, such as forums, or blogs are a good material for affect analysis because of their richness in evaluative and emotive information. One of that kind of services are blogs, open diaries in which people encapsulate their own experiences, opinions and feelings to be read and commented by other people. Recently blogs have come into the focus of scientific fields such as opinion mining, or sentiment and affect analysis [6]. Therefore creating a large blog-based corpus could become a solution to overcome both problems, of the lack in quantity of corpora and applicability of corpora in the research on emotions. However, to the authors' best knowledge, there has been only one small (4,186 sentences) Japanese blog corpus developed so far [2] (even though it was justified by the fact that corpus had to be manually annotated). Apart from that there exists another web-based corpus (JpWaC, [4]), however, although being considerably large (about 12 million sentences) it may still be insufficient for tasks, such as commonsense reasoning or context processing.

Therefore there was a need for a large-scale blog corpus able to be queried locally (e.g., when looking for word or sentence patterns), instead of querying the Internet with the use of search engines. Although there exist large resources, like Google N-gram Corpus [7], the textual data sets in such resources are short (up to 7-grams). This makes them unsuitable for emotion processing research, since most of contextual information, so important in expressing emotions [8] is lost.

We decided to create a new large-scale blog-based corpus from the scratch. The raw texts obtained from the web were not processed in any way (with exception to sentence segmentation). We aimed at creating a corpus that could serve as a source of various statistical data as well as a base for creating more detailed sub-corpora. Therefore we did not consider any pre-processing or additional tagging. Furthermore, on the base of this corpus we developed a sub-corpus containing emotion objects. Sentences in which emotions are expressed were extracted from the main corpus and annotated with emotion types. Particular markers were put to distinguish emotion object from emotional expression and the rest of the sentence. Similar project was already done for Chinese by Quan et al. [9], however no such research was done so far for Japanese. The corpus presented in this research is about 1000 times larger than Quan's, and in our research we used automatic annotation, while Quan's corpus was annotated manually. Other existing Japanese corpora designed for emotion research are small (e.g., the corpus by Minato et al. [10] contained only 1,200 sentences), which makes them unsuitable for most applications.

In the following sections we describe in details the tools and procedures used for compiling both, the main blog corpus and the corpus containing emotion objects. We present the detailed data about the both corpora, and propose possible applications, methods of improvement and future development of the corpora.

2. Yacis Corpus

The corpus (named Yacis Corpus, or Yet Another Corpus of Internet Sentences) was assembled using data obtained automatically from the pages of Ameba Blog (www.ameblo.co.jp, below referred to as Ameblo). There were two main reasons for using Ameblo. Firstly, the users are mostly Japanese so the risk that the links may lead to pages written in a language other than Japanese is small. Secondly, Ameblo has a clear structure of HTML source code, which makes it easy to extract only posts and comments omitting the irrelevant contents, such as advertisements or menu links.

All the tools used for compiling this corpus were developed especially for the purpose of this research. Although we tried several existing solutions, all of them were insufficient for our needs. All our tools were written in C# and are operating under MS Windows systems.

2.1 Corpus Compilation

We developed a simple but efficient web crawler designed to crawl exclusively Ameblo Web pages. The only pages taken into account were those containing Japanese posts (pages with legal disclaimers, as well as posts written in English were omitted). Initially we fed the crawler with 1000 links taken from Google (response to a query: 'site:ameblo.jp'). All the pages were saved to disk as raw HTML files (each page in a separate file) to be processed later. All of them were downloaded within 3 weeks between 3rd and 24th of December 2009. Next, we extracted all the posts and comments and divided them into sentences.

Although sentence segmentation may seem to be a trivial task it is not that easy when it comes to texts written by bloggers. People often use improper punctuation, e.g., the periods at the end of sentences are very often missing. In that case we assumed that if the given parts of text are separated by 2
 tags (two markers of a new line) then those parts will be two separate sentences. This does not solve the problem in all cases. Therefore we rejoined previously separated parts if the first part ended with a coma or if the quotation marks or parenthesis were opened in the first part and closed in second.

Unfortunately, these modifications were still not perfect and in several cases parts of the text remained not split while others were segmented erroneously. One of the possible improvements was to take into consideration emoticons. We observed that if an emoticon is present in the sentence it usually appears at the end of it. Using our emoticon analysis system [11] we intend to segment some more sentences. This method is currently being implemented and will be available in the second version of the corpus.

Currently the data is stored in modified-XML format. Although it looks like XML it does not comply with all XML standards due to the presence of some characters forbidden by XML specification, such as apostrophes (') or quotation marks ("). Those modifications were made to improve the communication with natural language processing tools that may be used in further processing of the corpus, such as a text parser, part-of-speech analyzer (e.g., JUMAN [12], MeCab [13]), affect analysis system (ML-ask [14]) and others. Each page was transformed into independent XML block between <doc></doc> tags. Opening tag of the <doc> block contains three parameters: URL, TIME and ID which specify the exact address from which the given page was downloaded, download time and unique page number, respectively. The <doc> block contains two other tag types: <post> and <comments>. The <post> block contains all the sentences from the given post where each sentence is included between <s></s> tags. The block <comments> contains all comments written under given post placed between <cmt></cmt> tags which are further split into single sentences placed between <s></s> tags (as described above).

The corpus is stored in 129 text files containing 100 000 <doc> units each. The corpus was encoded using UTF-8 encoding. The size of each file varies and is between 200 and 320 megabytes. The size of raw corpus (the corpus without any tags) is 27.1 gigabytes. Other statistics of the corpus are represented in the table below.

Table 1 Statistics of Yacis Corpus

# of sentences	354 288 529
# of web pages	12 938 606
# of unique bloggers	60 658
average # of pages/blogger	213.3
# of pages with comments	6 421 577
# of comments	50 560 024
average # of comment/page	7.873

As mentioned in the Table 1, average sentence length is 28.17 Japanese characters. Kubota et al. [15] divide sentences in Japanese according to their intelligibility into: easily intelligible short sentences (up to 100 characters) and difficult long sentences (over 100 characters long). The sentences in our corpus fit in the definition of short sentences which means they are easily understandable. After exclusion of very long sentences (consisting of over 500 characters) the number of sentences does not change significantly and is 354 169 311 (99,96%) with an average length of 27.9 characters. This means the corpus is balanced in the length of sentences.

2.2 Copyrights

Erjavec and colleagues claim that downloading texts and presenting various statistics obtained from them is an activity similar to those performed by search engines

[4]. Therefore there is no need for asking copyright holders for permission to use their texts in the corpus. The corpus is meant to be used for pure scientific purposes and is not planned to be available on sale. However, we wish to contribute with it to other Natural Language Processing research. Therefore we are open to make the corpus available to other researchers after specifying applicable legal conditions.

3. Database of Emotion Objects

The main purpose of this paper is to present Yacis Corpus. However, to demonstrate usefulness of the corpus in creating more specified sub-corpora, we created a sub-corpus of emotion objects, INFOE, on the basis of Yacis Corpus. An object of emotions was defined as "axiological properties which individuate emotions, make them intelligible and give them correctness conditions" [19]. In practice this means that emotion objects represent the context of one's expression of a given emotion. For example, in a sentence "I am afraid of spiders", the "spider" would be the emotion object of the emotion of "fear".

3.1 Compiling INFOE

We focused on emotions expressed in online communication in Japanese. Therefore, we needed to choose a classification of emotions proven to be the most appropriate for the Japanese language. We applied the general definition of emotions as every temporary state of mind, feeling, or affective state evoked by experiencing different sensations [20,21]. As for the classification of emotions, we applied that of Nakamura [21], who after over 30 years of thorough study in the lexicography of the Japanese language and emotive expressions, distinguishes 10 emotion types as the most appropriate for the Japanese language and culture. These are: ki/yorokobi (joy, delight), do/ikari (anger), ai/aware (sadness, gloom), fu/kowagari (fear), chi/haji (shame, shyness), ko/suki (liking, fondness), en/iya (dislike), ko/takaburi (excitement), an/yasuragi (relief) and kyo/odoroki (surprise, amazement). Nakamura uses this classification in his Dictionary of Emotive Expressions. The dictionary contains over two thousand expressions, each associated with one or more emotion type. Nakamura's emotion classification [21] and his dictionary of emotive expressions were used to prepare the corpus.

First we extracted from Yacis Corpus those sentences which contained at least one emotive expression. From the results we extracted sentences containing causality morpheme. Causality morphemes in Japanese are, e.g., the following: *-te, -to, -kara, -node, -tara*. Then, using a regular expression we extracted sentences in which causality expression is followed by emotive expression and distance between both is no greater than 5 characters. Both expressions were then tagged with proper markers. Also every sentence in the corpus containing emotive expressions was tagged with respective emotion type. All extraction and tagging was done automatically.

3.2 Semantic Formalization

We used *Bunrui Goihyo* [18] – word list containing semantic categories of words for semantic formalization. Each word in this list is associated with category (such as "abstract objects", "human activities", subject of actions, etc.) and one or more subcategories.

If the word from the list was found in the sentence a proper tag with a category number (unique category identifier) was added at the end of sentence. Very often more than one word from the sentence was found in the dictionary. Therefore tags occupy 60% of the size of the corpus.

4. Conclusions

In this paper we presented Yacis Corpus, a corpus of Japanese blogs compiled for the need of NLP research. We developed a set of tools for compilation of corpora and successfully compiled the large scale corpus from Ameblo web service. Although some work still needs to be done we believe that the main corpus containing over 300 million sentences is a valuable resource and could contribute greatly to numerous NLP research.

The eligibility of Yacis Corpus for the research on emotions in language has been already proved in our other research. We used it to evaluate our emoticon analysis system [11]. Also an emotion object database (INFOE) described in this paper was created based on Yacis Corpus proving it as sufficient base for creating more detailed subcorpora. Yacis Corpus can also be used in research on subjectivity or sentiment analysis, in which the use of the large corpora is becoming a standard [22]. Other possible applications are: research on common language use (statistical analysis), common sense and human creativity (psychology). After resolving problems with processing time (see below) the corpus may serve as an alternative for systems relying on constant search engine querying (e.g. chatbots).

5 Future Work

The size of all XML files in Yacis Corpus is 32.6 gigabytes. Reading through all of them takes some time (about 4 minutes on a computer with specs such as: Intel Core 2 Quad @ 2.4 GHz, 4GB RAM, Windows Vista Home Premium). This renders the corpus unsuitable for real time applications. Therefore the most desirable would be to apply indexing engine which would shorten the time of accessing the data. Furthermore, we plan to transfer the corpus into a database (MySql) to make the data accessible using SQL language. Finally, we intend to perform word segmentation and part-of-speech tagging using MeCab.

References

[1] Mainichi Shinbun CD, http://www.nichigai.co.jp/sales/mainichi/mainichi-data.html

[2] Kyoto University's NLP portal http://www-nagao.kuee.kyoto-u.ac.jp/NLP_Portal/lr-cat-e.html

[3] Aozora Bunko http://www.aozora.gr.jp/

[4] S. Erjavec, T. Erjavec, A. Kilgarriff, "A web corpus and word sketches for Japanese", *Technical Report*, 2008
[5] Ahmed Abbasi and Hsinchun Chen. "Affect Intensity Analysis of Dark Web Forums", *Intelligence and Security Informatics 2007*, pp. 282-288, 2007
[6] Shuya Abe, Moe Eguchi, Asuka Sumida, Azusa Ohsaki, Kentaro Inui. *"Minna no keiken: Burogu kara chuushutsu shita ibento oyobi senchimento no DB-ka"* [*Everyone's experiences: Creating a Database of Events and Sentiments Extracted from Blogs*] (in Japanese), In Proceedings of NLP-2009, pp. 296-299, 2009
[7] T. Kudo, H. Kazawa, *Japanese Web N-gram Version 1*, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T08
[8] Michal Ptaszynski, Rafal Rzepka and Kenji Araki, "On the Need for Context Processing in Affective Computing", In *Proceedings of Fuzzy System Symposium (FSS2010)*, Organized Session on Emotions, September 13-15, 2010 (to appear)
[9] Ch. Quan, F. Ren "Construction of a Blog Emotion Corpus for Chinese Emotional Expression Analysis", In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1446-1454, 2009.
[10] J. Minato, D. B. Bracewell, F. Ren, S. Kuroiwa "Japanese Emotion Corpus Analysis and its Use for Automatic Emotion Word Identification", *Engineering Letters*, 16:1, EL_16_1_25, 2009.
[11] Michal Ptaszynski, Jacek Maciejewski, Pawel Dybala, Rafal Rzepka and Kenji Araki, "CAO: Fully Automatic Emoticon Analysis System", In *Proc. of the 24th AAAI Conference on Artificial Intelligence (AAAI-10)*, pp. 1026-1032, 2010
[12] T. Kudo "MeCab: Yet Another Part of Speech and Morphological analyzer", http://mecab.sourceforge.net/
[13] Kyoto University, JUMAN 6.0 (a User-Extensible Morphological Analyzer for Japanese), http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/ juman-e.html, (2010.07.21)
[14] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka and Kenji Araki, "Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -", In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING-09)*, pp. 223-228, 2009.
[15] H. Kubota, K. Yamashita, T. Fukuhara, T. Nishida, "POC caster: Broadcasting Agent Using Conversational Representation for Internet Community" [in Japanese], *Transactions of the Japanese Society for Artificial Intelligence*, AI-17, pp. 313-321, 2002
[16] J. R. S. Wilson, "*Emotion and Object"*, Cambridge: Cambridge University Press, 1972
[17] K. Mulligan, "Intentionality, Knowledge and Formal Objects", *Electronic Festschrift for W. Rabinowicz*, T. Ronnow-Rasmussen et al. (eds.), 2007
[18] The National Institute for Japanese Language, "Bunrui Goihyo (Word List by Semantic Principles, Revised and Enlarged Edition)", http://www.kokken.go.jp/en/publications/bunrui_goihyo/
[19] F. Teroni, "Emotions and Formal Objects", *dialectica*, pp. 395-415, 2007.
[20] M. Lewis, J. M. Haviland-Jones, L. Feldman Barrett (eds.), "*Handbook of emotions"*, Guilford Press, 2008.
[21] Akira Nakamura, *"Kanjo hyogen jiten"* [*Dictionary of Emotive Expressions*] (in Japanese), Tokyodo Publishing, Tokyo, 1993.
[22] P. D. Turney, M. L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus", National Research Council of Canada, 2002.

Appendix 1

The example XML structure of the main blog corpus.

```
<doc url="http://ameblo.jp/capo-del-rosso/entry-000000.html" time="2009-12-05 21:11:46" id="2000001">
        <post>
                <s>今日から十月です。</s>
                [Its October from today.]
                <s>なんか、九月はいつもよりアッという間に過ぎたような気がするなぁ。</s>
                [I have a strange feeling September passed faster than usual.]
                ...
        </post>
        <comments>
                <cmt>
                        <s>色々と忙しいですね～！</s>
                        [Oh, you've been busy, weren't you?]
                        ...
                </cmt>
                <cmt>
                        <s>お疲れサマです(^O^)</s>
                        [Well done! Cheers for good work (^o^)]
                        ...
                </cmt>
        </comments>
</doc>
```

Appendix 2

The description of tags used in Yacis Corpus.

<doc>	- main tag, all the data retrieved from single website (both post and comments) is included between those tags
<post>	- contains all sentences from one post (included between <s></s> tags)
<comments>	- contains all comments written under given post (included between <cmt></cmt> tags)
<cmt>	- contains all sentences from single comment written under given post
<s>	- contains a single sentence (from either post or comment)