A Discriminative Approach to Japanese Abbreviation Extraction

Naoaki Okazaki[†]
okazaki@is.s.u-tokyo.ac.jp

Mitsuru Ishizuka[†]
ishizuka@i.u-tokyo.ac.jp

Jun'ichi Tsujii[†‡]
tsujii@is.s.u-tokyo.ac.jp

[†]Graduate School of Information
Science and Technology,
University of Tokyo
7-3-1 Hongo, Bunkyo-ku,
Tokyo 113-8656, Japan

[‡]School of Computer Science,
University of Manchester
National Centre for Text Mining (NaCTeM)
Manchester Interdisciplinary Biocentre,
131 Princess Street, Manchester M1 7DN, UK

Abstract

This paper addresses the difficulties in recognizing Japanese abbreviations through the use of previous approaches, examining actual usages of parenthetical expressions in newspaper articles. In order to bridge the gap between Japanese abbreviations and their full forms, we present a discriminative approach to abbreviation recognition. More specifically, we formalize the abbreviation recognition task as a binary classification problem in which a classifier determines a positive (abbreviation) or negative (non-abbreviation) class, given a candidate of abbreviation definition. The proposed method achieved 95.7% accuracy, 90.0% precision, and 87.6% recall on the evaluation corpus containing 7,887 (1,430 abbreviations and 6,457 non-abbreviation) instances of parenthetical expressions.

1 Introduction

Human languages are rich enough to be able to express the same meaning through different diction; we may produce different sentences to convey the same information by choosing alternative words or syntactic structures. Lexical resources such as WordNet (Miller et al., 1990) enhance various NLP applications by recognizing a set of expressions referring to the same entity/concept. For example, text retrieval systems can associate a query with alternative words to find documents where the query is not obviously stated.

Abbreviations are among a highly productive type of term variants, which substitutes fully expanded terms with shortened term-forms. Most previous studies aimed at establishing associations between abbreviations and their full forms in English (Park and Byrd, 2001; Pakhomov, 2002; Schwartz and Hearst, 2003; Adar, 2004; Nadeau and Turney, 2005; Chang and Schütze, 2006; Okazaki and Ananiadou, 2006). Although researchers have proposed various approaches to solving abbreviation recognition through methods such as deterministic algorithm, scoring function, and machine learning, these studies rely on the phenomenon specific to English abbreviations: all letters in an abbreviation appear in its full form.

However, abbreviation phenomena are heavily dependent on languages. For example, the term *one-segment broadcasting* is usually abbreviated as *one-seg* in Japanese; English speakers may find this peculiar as the term is likely to be abbreviated as *1SB* or *OSB* in English. We show that letters do not provide useful clues for recognizing Japanese abbreviations in Section 2. Elaborating on the complexity of the generative processes for Japanese abbreviations, Section 3 presents a supervised learning approach to Japanese abbreviations. We then evaluate the proposed method on a test corpus from newspaper articles in Section 4 and conclude this paper.

2 Japanese Abbreviation Survey

Researchers have proposed several approaches to abbreviation recognition for non-alphabetical languages. Hisamitsu and Niwa (2001) compared different statistical measures (e.g., χ^2 test, log like-

Type	Para	#	(%)	Examples	
Acronym	o	90	(1.2)	東京大学（東大） Tokyo Daigaku (ToDai) The University of Tokyo (UoT)	首都圏中央連絡自動車道（圏央道） Shutoken Chuou Renraku Jidousha Dou (Ken-o-dou) Metropolitan Inter-City Expressway (Ken-o Exp)
Acronym with translation	o	717	(9.1)	夜間離着陸訓練（ＮＬＰ） Yakan Richakuriku Kunren (NLP) Night Landing Practice (NLP)	ワールドカップ（Ｗ杯） Warudo Kappu (W hai) World Cup (WC)
Alias	o	623	(7.9)	朝鮮民主主義人民共和国（北朝鮮） Cho-sen Minshushugi Jinmin Kyowakoku (Kita Chosen) Democratic People's Republic of Korea (North Korea)	2000年問題（Ｙ２Ｋ） 2000 Nen Mondai (Y2K) Year 2000 problem (Y2K)
Attribute (reading)	x			毅然（きぜん） Kizen (kizen) firm [fəːrm]	Ｏ（オー）１５７ O (O-) 157
Attribute (location)	x			つくば学園都市（茨城県つくば市） Tsukuba Gakuen Toshi (Ibaraki-ken Tsukuba-shi) Tsukuba Science City (Tukuba City, Ibaraki Pref.)	
Attribute (affiliation)	x	6,457	(81.9)	インディペンデント（英国） Indipendento (Eikoku) Independent (UK)	ミヒャエル・シューマッハー（独） Mihyaeru Shumahha- (GER) Michael Schumacher (GER)
Epexegesis	x			西ドイツ（当時） Nishi Doitsu (touji) West Germany (at that time)	平成金融再生機構（仮称） Heisei Kinyu Saisei Kikou (Kasho) Financial Revitalization Corporation (tentative name)
Others	x			…（中略） ... (churyaku) ... (snip)	

Table 1: Parenthetical expressions used in Japanese newspaper articles

lihood ratio) to assess the co-occurrence strength between the inner and outer phrases of parenthetical expressions *X (Y)*. Yamamoto (2002) utilized the similarity of local contexts to measure the paraphrase likelihood of two expressions based on the distributional hypothesis (Harris, 1954). Chang and Teng (2006) formalized the generative processes of Chinese abbreviations with a noisy channel model. Sasano et al. (2007) designed rules about letter types and occurrence frequency to collect lexical paraphrases used for coreference resolution.

How are these approaches effective in recognizing Japanese abbreviation definitions? As a preliminary study, we examined abbreviations described in parenthetical expressions in Japanese newspaper articles. We used the 7,887 parenthetical expressions that occurred more than eight times in Japanese articles published by the *Mainichi Newspapers* and *Yomiuri Shimbun* in 1998–1999. Table 1 summarizes the usages of parenthetical expressions in four groups. The field 'para' indicates whether the inner and outer elements of parenthetical expressions are interchangeable.

The first group *acronym* (I) reduces a full form to a shorter form by removing letters. In general, the process of acronym generation is easily interpreted: the left example in Table 1 consists of two Kanji letters taken from the heads of the two words, while the right example consists of the letters at the end of the 1st, 2nd, and 4th words in the full form. Since all letters in an acronym appear in its full form, previous approaches to English abbreviations are also applicable to Japanese acronyms. Unfortunately, in this survey the number of such 'authentic' acronyms amount to as few as 90 (1.2%).

The second group *acronym with translation* (II) is characteristic of non-English languages. Full forms are imported from foreign terms (usually in English), but inherit the foreign abbreviations. The third group *alias* (III) presents generic paraphrases that cannot be interpreted as abbreviations. For example, *Democratic People's Republic of Korea* is known as its alias *North Korea*. Even though the formal name does not refer to the 'northern' part, the alias consists of *Korea*, and the locational modifier *North*. Although the second and third groups retain their interchangeability, computers cannot recognize abbreviations with their full forms based on letters.

The last group (IV) does not introduce interchangeable expressions, but presents additional information for outer phrases. For example, a location usage of a parenthetical expression *X (Y)* describes an entity *X*, followed by its location *Y*. Inner and outer elements of parenthetical expressions are not interchangeable. We regret to find that as many as 81.9% of parenthetical expressions were described for this usage. Thus, this study regards acronyms (with and without translation) and alias as *Japanese*

#	Expression *Y*	Expression *X*	Freq	Class
1	北朝鮮 (North Korea)	朝鮮民主主義人民共和国 (Democratic People's Republic of Korea)	4160	A
2	W杯 (W Cup)	ワールドカップ (World Cup)	2891	T
3	EU	欧州連合 (European Union)	2638	T
4	NATO	北大西洋条約機構 (North Atlantic Treaty Organization)	2593	T
5	IMF	国際通貨基金 (International Monetary Fund)	2473	T
6	中国 (China)	人民日報 (People's Daily)	1561	O
7	IOC	国際オリンピック委員会 (International Olympic Committee)	1550	T
8	WTO	世界貿易機関 (World Trade Organization)	1504	T
9	独 (Germany)	ディ・ウェルト (Die Welt)	1484	O
10	エジプト (Egypt)	アルアハラム (Al-Ahram)	1350	O

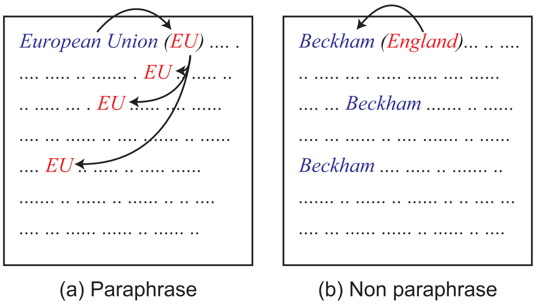
Table 2: Top 10 frequent parenthetical expressions used in Japanese newspapers from 1998–1999

abbreviations in a broad sense, based on their interchangeabilities. In other words, the goal of this study is to classify parenthetical expressions *X (Y)* into true abbreviations (groups I, II, III) and other usages of parentheses (group IV).

How much potential do statistical approaches have to identify Japanese abbreviations? Table 2 shows the top 10 most frequently appearing parenthetical expressions in this survey. The 'class' field represents the category[1]: *T: acronym with translation*, *A: alias*, and *O: non-abbreviation*. The most frequently occurring parenthetical expression was *Democratic People's Republic of Korea (North Korea)* (4,160 occurrences). 7 instances in the table were acronyms with translation (#2–5, #7–8), and an alias (#1), but 3 non-abbreviation instances (#6, #9, and #10) expressed nationalities of information sources. Even if we designed a simple method to choose the top 10 parenthetical expressions, the recognition performance would be no greater than 70% precision.

3 A discriminative approach to abbreviation recognition

In order to bridge the gap between Japanese abbreviations and their full forms, we present a discriminative approach to abbreviation recognition. More specifically, we formalize the abbreviation recognition task as a binary classification problem in which a classifier determines a positive (abbreviation) or negative (non-abbreviation) class, given a parenthetical expression *X (Y)*. We model the classifier by using Support Vector Machines (SVMs) (Vapnik, 1998). The classifier combines features that characterize various aspects of abbreviation definitions. Table 3 shows the features and their values for the abbreviation *EU*, and its full form: *O-shu Rengo (European Union)*. A string feature is converted into a set of boolean features, each of which indicates 'true' or 'false' of the value. Due to the space limitation, the rest of this section elaborates on *paraphrase ratio* and *SKEW* features.



(a) Paraphrase (b) Non paraphrase

Figure 1: Paraphrase occurrence with parentheses

Paraphrase ratio Let us consider the situation in which an author describes an abbreviation definition *X (Y)* to state a paraphrase *X* → *Y* in a document. The effect of the statement is to define the meaning of the abbreviation *Y* as *X* in case the reader may be unaware/uncertain of the abbreviation *Y*. For example, if an author wrote a parenthetical expression, *Multi-Document Summarization (MDS)*, in a document, readers would recognize the meaning of the expression *MDS*. Even if they were aware of the definition, *MDS* alone would be ambiguous; it could stand for *Multi Dimensional Scaling*, *Missile Defense System*, etc. Therefore, an author rarely uses the expression *Y* before describing its definition.

At the same time, the author would use the expression *Y* more than *X* after describing the definition, if it were to declare the abbreviation *Y* for *X*. Figure 1 illustrates this situation with two documents. Document (a) introduces the abbreviation *EU* for *European Union* because the expression *EU* occurs more frequently than *European Union* after the parenthetical expression. In contrast, the parenthetical expres-

[1] No *acronym* was included in the top 10 list.

Feature	Type	Description	Example
$\mathrm{PR}(X,Y)$	numeric	Paraphrase ratio	0.426
$\mathrm{SKEW}(X,Y)$	numeric	Similarity of local contexts measured by the skew divergence	1.35
$\mathrm{freq}(X)$	numeric	Frequency of occurrence of X	2,638
$\mathrm{freq}(Y)$	numeric	Frequency of occurrence of Y	8,326
$\mathrm{freq}(X,Y)$	numeric	Frequency of co-occurrence of X and Y	3,121
$\chi^2(X,Y)$	numeric	Co-occurrence strength measured by the χ^2 test	2,484,521
$\mathrm{LLR}(X,Y)$	numeric	Co-occurrence strength measured by the log-likelihood ratio	6.8
$\mathrm{match}(X,Y)$	boolean	Predicate to test whether X contains all letters in Y	0
Letter types	string	Pair of letter types of X and Y	Kanji/Alpha
First letter	string	The first letter in the abbreviation Y	*E*
Last letter	string	The last letter in the abbreviation Y	*U*
POS tags	string	Pair of POS tags for X and Y	NNP/NNP
POS categories	string	Pair of POS categories for X and Y	NN/NN
NE tags	string	Pair of NE tags for X and Y	ORG/ORG

Table 3: Features for the SVM classifier and their values for the abbreviation *EU*.

sion in document (b) describes the property (nationality) of a person *Beckham*.

Suppose that we have a document that has a parenthetical expression with expressions X and Y. We regard a document introducing an abbreviation Y for X if the document satisfies both of these conditions:

1. The expression Y appears more frequently than the expression X does after the definition pattern.

2. The expression Y does not appear before the definition pattern.

Formula 1 assesses the paraphrase ratio of the expressions *X* and *Y*,

$$\mathrm{PR}(X,Y) = \frac{d_{\mathrm{para}}(X,Y)}{d(X,Y)}. \quad (1)$$

In this formula, $d_{\mathrm{para}}(X,Y)$ denotes the number of documents satisfying the above conditions, and $d(X,Y)$ presents the number of documents having the parenthetical expression $X(Y)$. The function PR(X, Y) ranges from 0 (no abbreviation instance) to 1 (all parenthetical expressions introduce the abbreviation).

Similarity of local contexts We regard words that have dependency relations from/to the target expression as the local contexts of the expression, applying all sentences to a dependency parser (Kudo and Matsumoto, 2002). Collecting the local context of the target expressions, we compute the skew divergence (Lee, 2001), which is a weighted version of Kullback-Leibler (KL) divergence, to measure the resemblance of probability distributions P and Q:

$$\mathrm{SKEW}_{\alpha}(P||Q) = \mathrm{KL}(P||\alpha Q + (1-\alpha)P), \quad (2)$$

$$\mathrm{KL}(P||Q) = \sum_{i} P(i) \log \frac{P(i)}{Q(i)}. \quad (3)$$

In these formulas, P is the probability distribution function of the words in the local context for the expression X, Q is for Y, and α is a skew parameter set to 0.99. The function $\mathrm{SKEW}_{\alpha}(P||Q)$ becomes close to zero if the probability distributions of local contexts for the expressions X and Y are similar.

Other features In addition, we designed twelve features for abbreviation recognition: five features, $\mathrm{freq}(X)$, $\mathrm{freq}(Y)$, $\mathrm{freq}(X,Y)$, $\chi^2(X,Y)$, and $\mathrm{LLR}(X,Y)$ to measure the co-occurrence strength of the expressions X and Y (Hisamitsu and Niwa, 2001), $\mathrm{match}(X,Y)$ feature to test whether or not all letters in an abbreviation appear in its full form, three features *letter type*, *first letter*, and *last letter* corresponding to rules about letter types in abbreviation definitions, and three features *POS tags*, *POS categories*, and *NE tags* to utilize information from a morphological analyzer and named-entity tagger (Kudo and Matsumoto, 2002).

4 Evaluation

4.1 Results

We built a system for Japanese abbreviation recognition by using the LIBSVM implementation[2] with a

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvm

Group	Recall
Acronym	94.4%
Acronym with translation	97.4%
Alias	81.4%
Total	87.6%

Table 4: Recall for each role of parentheses

linear kernel, which obtained the best result through experiments. The performance was measured under a ten-fold cross-validation on the corpus built in the survey, which contains 1,430 abbreviation instances and 6,457 non-abbreviation instances.

The proposed method achieved 95.7% accuracy, 90.0% precision, and 87.6% recall for recognizing Japanese abbreviations. We cannot compare this performance directly with the previous work because of the differences in the task design and corpus. For reference, Yamamoto (2002) reported 66% precision (he did not provide the recall value) for a similar task: the acquisition of lexical paraphrase from Japanese newspaper articles.

Table 4 reports the recall value for each group of abbreviations. This analysis shows the distribution of abbreviations unrecognized by the proposed method. Japanese acronyms, acronyms with translation, and aliases were recognized at 94.4%, 97.4%, and 81.4% recall respectively. It is interesting to see that the proposed method could extract acronyms with translation and aliases even though we did not use any bilingual dictionaries.

4.2 Analyses for individual features

The numerical and boolean features are monotone increasing functions (decreasing for the SKEW feature) as two expressions X and Y are more likely to present an abbreviation definition. For example, the more authors introduce a paraphrase $X \rightarrow Y$, the higher the value that $\mathrm{PR}(X,Y)$ feature yields. Thus, we emulate a simple classifier for each feature that labels a candidate of abbreviation definition as a positive instance only if the feature value is higher than a given threshold θ, e.g., $\mathrm{PR}(X,Y) > 0.9$. Figure 2 shows the precision–recall curve for each feature with variable thresholds.

The paraphrase ratio (PR) feature outperformed other features with a wide margin: the precision and recall values for the best F1 score were 66.2% and 48.1% respectively. Although the performance of this feature alone was far inferior to the proposed method, to some extent Formula 1 estimated actual occurrences of abbreviation definitions.

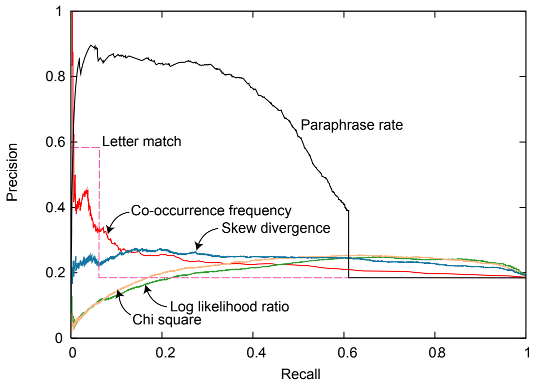


Figure 2: Precision–recall curve of each feature

Feature	Accuracy	Reduction
All	95.7%	—
- $\mathrm{PR}(X,Y)$	95.2%	0.5%
- $\mathrm{SKEW}(X,Y)$	95.4%	0.3%
- $\mathrm{freq}(X,Y)$	95.6%	0.1%
- $\chi^2(X,Y)$	95.6%	0.1%
- $\mathrm{LLR}(X,Y)$	95.3%	0.4%
- $\mathrm{match}(X,Y)$	95.5%	0.2%
- Letter type	94.5%	1.2%
- POS tags	95.6%	0.1%
- NE tags	95.7%	0.0%

Table 5: Contribution of the features

The performance of the match (letter inclusion) feature was as low as 58.2% precision and 6.9% recall[3]. It is not surprising that the match feature had quite a low recall, because of the ratio of 'authentic' acronyms (about 6%) in the corpus. However, the match feature did not gain a good precision either. Examining false cases, we found that this feature could not discriminate cases where an outer element contains its inner element accidentally; e.g., *Tokyo Daigaku (Tokyo)*, which describes a university name followed by its location (prefecture) name.

Finally, we examined the contribution of each feature by eliminating a feature one by one. If a feature was important for recognizing abbreviations, the absence of the feature would drop the accuracy. Each row in Table 5 presents an eliminated feature, the accuracy without the feature, and the reduction of

[3] This feature drew the precision–recall locus in a stepping shape because of its discrete values (0 or 1).

the accuracy. Unfortunately, the accuracy reductions were so few that we could not discuss contributions of features with statistical significance. The letter type feature had the largest influence (1.2%) on the recognition task, followed by the paraphrase ratio (0.5%) and log likelihood ratio (0.4%).

5 Conclusion

In this paper we addressed the difficulties in recognizing Japanese abbreviations by examining actual usages of parenthetical expressions in newspaper articles. We also presented the discriminative approach to Japanese abbreviation recognition, which achieved 95.7% accuracy, 90.0% precision, and 87.6% recall on the evaluation corpus. A future direction of this study would be to apply the proposed method to other non-alphabetical languages, which may have similar difficulties in modeling the generative process of abbreviations. We also plan to extend this approach to the Web documents.

Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (MEXT, Japan), and Solution-Oriented Research for Science and Technology (JST, Japan). We used *Mainichi Shinbun* and *Yomiuri Shinbun* newspaper articles for the evaluation corpus.

References

Eytan Adar. 2004. SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics*, 20(4):527–533.

Jeffrey T. Chang and Hinrich Schütze. 2006. Abbreviations in biomedical text. In S. Ananiadou and J. McNaught, editors, *Text Mining for Biology and Biomedicine*, pages 99–119. Artech House, Inc.

Jing-Shin Chang and Wei-Lun Teng. 2006. Mining atomic chinese abbreviation pairs: A probabilistic model for single character word recovery. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, pages 17–24, Sydney, Australia, July. Association for Computational Linguistics.

Zellig S. Harris. 1954. Distributional structure. *Word*, 10:146–162.

Toru Hisamitsu and Yoshiki Niwa. 2001. Extracting useful terms from parenthetical expression by combining simple rules and statistical measures: A comparative evaluation of bigram statistics. In Didier Bourigault, Christian Jacquemin, and Marie-C L'Homme, editors, *Recent Advances in Computational Terminology*, pages 209–224. John Benjamins.

Taku Kudo and Yuji Matsumoto. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the CoNLL 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69.

Lillian Lee. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics 2001*, pages 65–72.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):235–244.

David Nadeau and Peter D. Turney. 2005. A supervised learning approach to acronym identification. In *8th Canadian Conference on Artificial Intelligence (AI'2005) (LNAI 3501)*, pages 319–329.

Naoaki Okazaki and Sophia Ananiadou. 2006. A term recognition approach to acronym recognition. In *Proceedings of the COLING-ACL 2006 Main Conference Poster Sessions*, pages 643–650, Sydney, Australia.

Serguei Pakhomov. 2002. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of 40th annual meeting of ACL*, pages 160–167.

Youngja Park and Roy J. Byrd. 2001. Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the EMNLP 2001*, pages 126–133.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2007. Improving coreference resolution using bridging reference resolution and automatically acquired synonyms. In *Anaphora: Analysis, Alogorithms and Applications, 6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC2007*, pages 125–136.

Ariel S. Schwartz and Marti A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing (PSB 2003)*, number 8, pages 451–462.

Vladimir N. Vapnik. 1998. *Statistical Learning Theory*. John Wiley & Sons.

Kazuhide Yamamoto. 2002. Acquisition of lexical paraphrases from texts. In *2nd International Workshop on Computational Terminology (Computerm 2002, in conjunction with COLING 2002)*, pages 1–7, Morristown, NJ, USA. Association for Computational Linguistics.