An Easily Implemented Method for Abbreviation Expansion for the Medical Domain in Japanese Text

A Preliminary Study

E. Y. Shinohara[1]; E. Aramaki[2]; T. Imai[3]; Y. Miura[4]; M. Tonoike[4]; T. Ohkuma[4]; H. Masuichi[4]; K. Ohe[1,5]

[1]Department of Planning, Information and Management, The University of Tokyo Hospital, Tokyo, Japan;
[2]Center for Knowledge Structuring, The University of Tokyo, Tokyo, Japan;
[3]Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan;
[4]Research & Technology Group, Fuji Xerox Co., Ltd, Kanagawa, Japan;
[5]Graduate School of Medicine and Faculty of Medicine, The University of Tokyo, Tokyo, Japan

Keywords
Natural language processing, machine learning, abbreviation, information storage and retrieval, algorithms

Summary

Background: One of the barriers for the effective use of computerized health-care related text is the ambiguity of abbreviations. To date, the task of disambiguating abbreviations has been treated as a classification task based on surrounding words. Application of this framework for languages that have no word boundaries requires pre-processing to segment a sentence into separate word sequences. While the segmentation processing is often a source of problem, it is unknown whether word information is really requisite for abbreviation expansion.

Objectives: The present study examined and compared abbreviation expansion methods with and without the incorporation of word information as a preliminary study.

Methods: We implemented two abbreviation expansion methods: 1) a morpheme-based method that relied on word information and therefore required pre-processing, and 2) a character-based method that relied on simple character information. We compared the expansion accuracies for these two methods using eight medical abbreviations. Experimental data were automatically built as a pseudo-annotated corpus using the Internet.

Results: As a result of the experiment, accuracies for the character-based method were from 0.890 to 0.942 while accuracies for the morpheme-based method were from 0.796 to 0.932. The character-based method significantly outperformed the morpheme-based method for three of the eight abbreviations ($p < 0.05$). For the remaining five abbreviations, no significant differences were found between the two methods.

Conclusions: Character information may be a good alternative in terms of simplicity to morphological information for abbreviation expansion in English medical abbreviations appeared in Japanese texts on the Internet.

Correspondence to:
Emiko Yamada Shinohara
7-3-1 Hongo, Bunkyo-ku
Tokyo 113-8655
Japan
E-mail: emiko-tky@umin.net

Methods Inf Med 2013; 52: 51–61
doi: 10.3414/ME12-01-0040
received: April 30, 2012
accepted: October 28, 2012
prepublished: December 7, 2012

1.Introduction

The prevalence of computerized texts within the health care domain is increasing with the widespread adoption of hospital information systems. Such data might be used for many promising applications, including detection of adverse drug events, support decision making, and surveillance [1–3].

In actuality, however, synonym and multi-sense words are obstacles for such applications [4]. If one searches through clinical notes using the search query "diabetes mellitus" in order to discern how many patients with diabetes mellitus visited the hospital, they may obtain results that are unreliable, because they may miss documents that use only the abbreviation "DM". Expanding the search query to include "diabetes mellitus or DM" would solve this problem, but would create a new problem: the results for this query may become very noisy, and may contain other possible expansions of "DM", such as, "dermal melanosis", "dermatomyositis", "diffuse, mixed, small and large cell" and "dystrophia myotonica".

As a result, as abbreviations are ambiguous, it is often unclear to which full form an abbreviation corresponds. Not only information retrieval, but various other applications, such as speech recognition [5] and information extraction [6], also suffer from complications related to abbreviations. Moreover, ambiguities in abbreviations appearing in clinical notes may lead

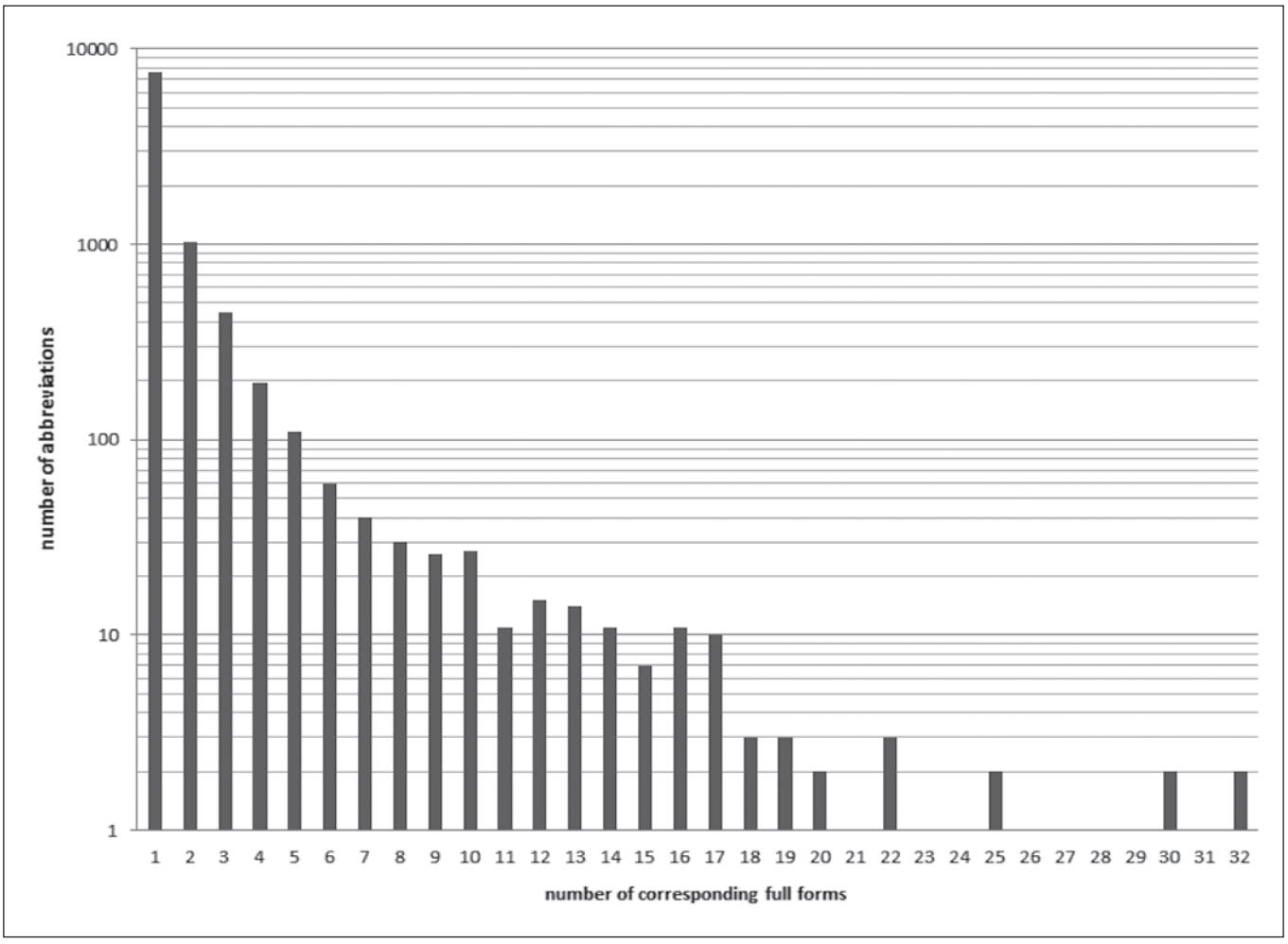

Figure 1 Ambiguity of Japanese medical abbreviation. This figure presents the number of abbreviations with respect to each of their ambiguity. While the most abbreviations' possible full forms are just one, there are numerous ambiguous abbreviations.

to misunderstandings and errors in clinical practice. In order to resolve this problem, the Joint Commission on Accreditation of Healthcare Organizations in the US has issued a "Do Not Use" list for abbreviations that presents the most egregious and dangerous abbreviations where patient care errors can potentially occur, although compliance with this list may be insufficient [7].

One potential approach that may decrease ambiguity is to prevent the use of abbreviations in clinical notes. For example, in 2007 Myers et al. examined an alert system that would prevent users from inputting unapproved abbreviations into medical records and found that the system significantly reduced the use of ambiguous abbreviations [8]. However, large amounts of text containing abbreviations are already in existence, and it is difficult to install functioning alert systems in existing data input systems. This situation motivated us to devise a way of disambiguating abbreviations automatically – specifically, a method to expand abbreviations into their full form.

For global abbreviations that appear in documents without the full form explicitly stated [9], abbreviation expansion is generally treated as a classification task – that is, selecting a full form from a list of full forms for a target abbreviation. This is technically similar to "word sense disambiguation" (WSD) in which full forms are regarded as word senses. WSD is a task which selects one meaning from sense inventory given an ambiguous word and context, e.g., given "bank" as the ambiguous word and "I withdrew some money from the bank" as the context, selecting one of the senses from the inventory "financial bank" and "river bank". The study of WSD has a long history in the field of natural language processing (NLP). Although various methods have been proposed for both abbreviation expansion and WSD, most researchers have adopted a supervised approach [17–29], while others have adopted an unsupervised approach [30–32]. For global abbreviation expansion, Joshi et al. compared supervised machine learning algorithms and feature sets [25]. While there are no standard full form lists that cover abbreviations that appear in clinical texts [10, 11], several researchers have attempted to compile (semi-)automatic lists utilizing local abbreviations, where the full form is explicitly stated, as well as resources such as UMLS [11–16]. Okazaki et al. proposed an

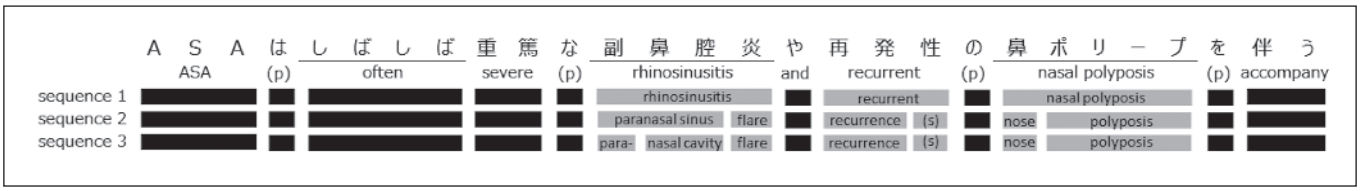

Figure 2 Morphemes have inherent ambiguity. Possible morpheme sequences of the sentence "ASA are often accompanied by severe rhinosinusitis and recurrent nasal polyposis" in Japanese. (p) denotes a Japanese particle that indicates the grammatical role of its preceding word. (s) denotes suffixes, some of which change the part of speech of its preceding word. Note that "rhinosinusitis", "recurrent" and "nasal polyposis" contain ambiguity in terms of morpheme sequences.

integrative method to expand global abbreviations using local abbreviations that appear in the same corpus [33]. The present paper focuses on global abbreviations.

The meaning of each abbreviation is fundamentally determined by its context, which in most cases is contingent upon the surrounding words. The following are examples of ambiguous text associated with the abbreviation **ASA**:

1. Antiplatelet action appears after administration of low dose **ASA**.
2. **ASA** is often accompanied by severe rhinosinusitis and recurrent nasal polyposis.

The meaning of ASA differs between the two sentences. According to a medical abbreviation dictionary [34], ASA is an abbreviation of five different full forms in the medical domain: i) "acetylsalicylic acid", a type of medicine, ii) "active systemic anaphylaxis", a type of disease, iii) "anti-smooth muscle antibody", a type of chemical compound, iv) "argininosuccinic acid", a type of chemical compound, and v) "aspirin sensitive asthma", a type of disease. In the first example, the context "administration of low dose" suggests that the ASA is a medicine that immediately determines the full form "acetylsalicylic acid". The second example refers to a disease, which can be inferred from the phrases "accompanied by severe rhinosinusitis." Therefore, the candidates for the 2nd ASA are "active systemic anaphylaxis" and "aspirin-sensitive asthma." The final clues may be "Severe rhinosinusitis" and "recurrent nasal" which help decipher the correct full form "aspirin-sensitive asthma."

To date, most studies have targeted English abbreviations in English text [23–28]. More recently, English abbreviations in other languages have also been examined [29]. One medical abbreviation dictionary in Japanese contains 2022 ambiguous English abbreviations, each corresponding to up to 32 full forms, and 3.7 full forms on average (▶Figure 1) [34]. For clinical texts, we investigated a small selection of Japanese discharge summaries and found that one discharge summary contained approximately 13 abbreviations on average and each abbreviation corresponded to 1–2 full forms. Thus, we believe that the „actual" ambiguity is much greater. The present study specifically examined English abbreviations in Japanese text. A typical Japanese sentence uses Latin alphabet characters along with Japanese characters. In most cases, sequences of Latin alphabet characters in Japanese text represent English abbreviations. For example, the second sentence in the example provided above is represented in Japanese as shown in ▶Figure 2.

In order to expand the abbreviation ASA, the same methods apply as in English text. However, one extra pre-processing step is necessary to split the sentence into word sequences to capture the surrounding words that are used as clues for expansion. One difference between English and Japanese when NLP is to be applied is the use of word boundaries: Japanese has no explicit word boundaries, which necessitates a word segmentation step. In practice, word segmentation is done through morphological analysis, which splits a sentence into a sequence of morphemes – that is, the smallest semantically meaningful unit, approximately corresponding to a word in English. However, this approach suffers from analytic errors (▶Figure 3). This is because a general morphological analyzer is designed to analyze newspaper text. This characteristic is a general barrier in processing Japanese text [35].

Furthermore, the definition of a morpheme is ambiguous in a fundamental sense. The key problem here is how to identify the best representation of context, in order to determine the classification for the abbreviation that precedes or follows. Figure 2 presents a few possible morpheme sequences for the ASA example. The sequences differ in their granularity. The first sequence represents the finest grain. The more the sentence is split, the more general

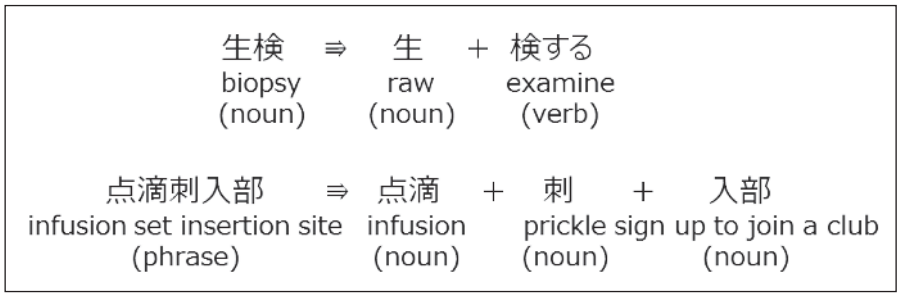

Figure 3 Error of morphological analysis. A morphological analyzer sometimes splits a morpheme that should not be split, e.g., "Biopsy". An additional type of error is that the analyzer may split a sentence at the wrong position. For example, the following phrase "Infusion set insertion site" should be split into three morphemes, "infusion", "insertion", and "site".

the meaning of the word. Thus far, no study of optimal granularity for abbreviation expansions has been reported, perhaps due to the fact that the answer depends on the target abbreviation. "Rhinosinusitis" may be a clue to expand certain abbreviations, while for others "paranasal sinus flare" may contain key terms rather than the entire disease name. To ascertain a suitable morpheme sequence for this specific purpose, one must compile a dictionary for use by a morphological analyzer, along with many parameters, e.g., connective costs for pairs of morphemes/part of speech (POS), for each abbreviation. This situation motivated us to skip morphological analysis.

Thus, we present here an abbreviation expansion method based on character contexts, not morpheme contexts. Several reports have demonstrated that a character-based approach is superior to a morpheme-based approach in tasks such as WSD [36], information retrieval [37], and translation memory for machine translation [38].

In addition to Japanese, Korean and Chinese languages entail similar difficulties. Although the present study specifically addresses Japanese, the proposed method is expected to be applicable to other languages as well because the method does not need any language-specific processing.

Although abbreviation expansion is one of the critical steps for clinical text processing, it is very difficult to construct a corpus from clinical text. Therefore, the present study examined the possibility of character based abbreviation expansions using text from documents on the World Wide Web as a preliminary study.

Research question: Is abbreviation expansion with morphological analysis truly superior to abbreviation expansion without morphological analysis?

2. Materials and Methods

To address this research question, we implemented two abbreviation expansion methods and compared them. Specifically, we compared the two expansion methods with and without morphological analysis. This section presents a description of these two methods, materials, and evaluation scale.

2.1 Abbreviation Expansion Methods

As in most previous studies where the target was global abbreviation, we treated abbreviation expansion as a classification of an abbreviation into one full form within a given full form list based on its context. An abbreviation corresponds to one or more full forms. Therefore, this task is a multi-class problem. The present study adopted a one-versus-the-rest classifier, which used a collection of binary classifiers, each of which corresponded to one full form and classified input into either the full form or "the rest" category (▶Figure 4). We used a support vector machine (SVM) known for its high generalization ability and high performance for abbreviation expansion [25] as a binary classifier.

The two methods compared share the classification framework described above, although they differ in the usage of context on which the expansion was done.

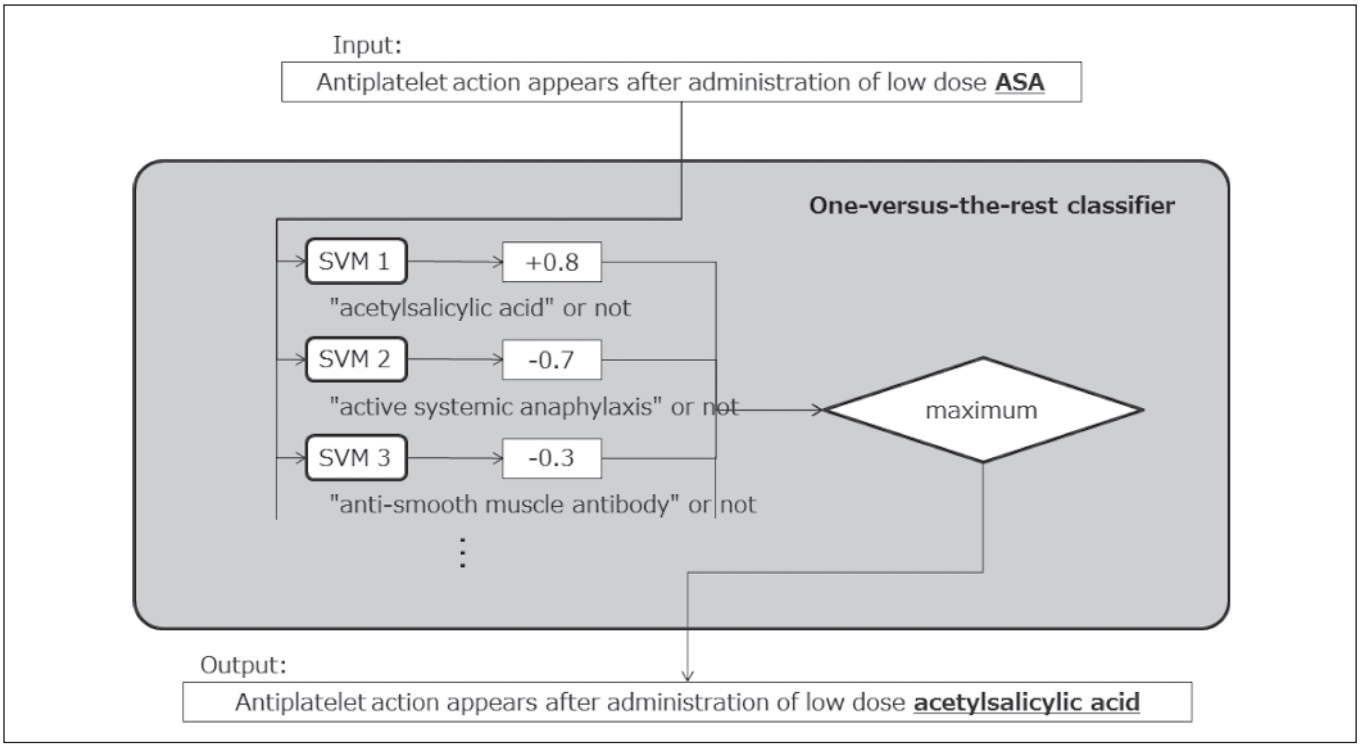

Figure 4 One-versus-the-rest classifier. In this figure, SVM output numbers represent distances from the separating hyperplanes, not binary judgment.

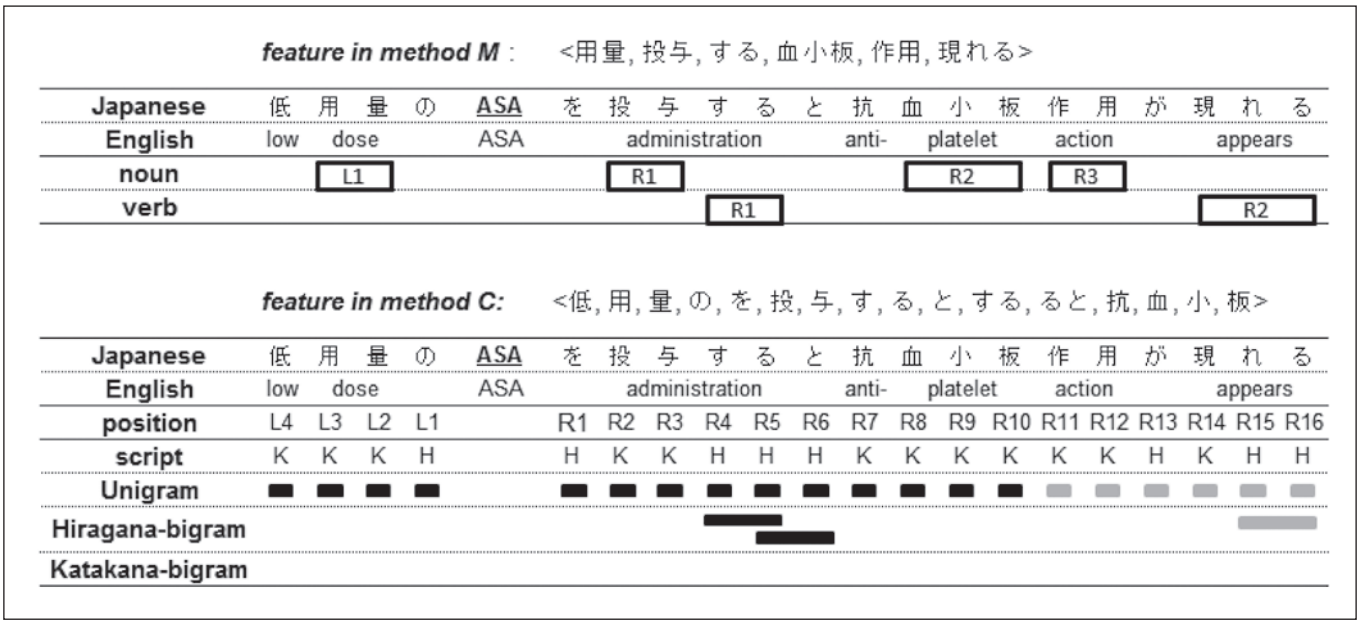

Figure 5 Feature example by two methods (***Method M*** and ***Method C***). This figure illustrates the features of the sentence "Antiplatelet action appears after administration of low dose ASA" by the two methods. ***Method M*** uses 6 words as features. Li and Ri indicate the i-th words on the left and right, respectively. ***Method C*** (window size: 10) uses 14 unigrams (9 kanji and 5 hiragana), 2 hiragana-bigrams and no katakana-bigrams (total 16 n-grams). Li and Ri indicate the relative position to the abbreviation. "script" indicates script-type, K for kanji and H for hiragana; katakana was not included. Black bars show corresponding strings contained in the feature.

2.1.1 Morpheme-based Method (Method M)

The first method expands abbreviations based on nearby morphemes. As in our previous study [29], we used verbs and nouns among morphemes: three verbs and three nouns prior/posterior to the target abbreviation (with a maximum 12 morphemes total). This is a simplified version of a technique used in a previous study [25], in which all words were used except function words and stop words defined by the authors (e.g. proper noun). In order to avoid listing appropriate stop words in Japanese, we simplified their method and ruled out adjectives and adverbs. Because adjectives are often made synthetically from a noun and a suffix in Japanese medical text (as shown in ▶Figure 3: adjective "recurrent" is split into the noun "recurrence" and a suffix.) and adverbs are thought to be less informative, the simplification would not disadvantage a morpheme-based method. The window size was also decided based on the previous study [25], which demonstrated that a window size of three worked well. Thus, we used MeCab[a] as the Japanese morphological analyzer and UniDic[b] as a dictionary for MeCab. MeCab is a conditional random field (CRF)-based high-performance morphological analyzer. UniDic is a uniformly designed Japanese dictionary for the general domain.

2.1.2 Character-based Method (Method C)

The character-based method expands abbreviations based on nearby characters. More precisely, an abbreviation is classified into one full form based on prior and subsequent n characters (n is referred to as "window size" in the present study).

The Japanese language uses three types of written characters: 1) hiragana, 2) katakana, and 3) kanji (Chinese characters). Although 1) hiragana and 2) katakana are phoneme-based, 3) kanji are ideograms for which a single character indicates a concept. For example, one kanji may mean "bone", while another means "disease". To even out the information per feature, we adopted "characters, hiragana bigrams, and katakana bigrams" as a feature set (▶Figure 5).

These two aforementioned methods were compared. In order to balance the information presented in the two methods, the character-based method window size was set to 20 because a Japanese morpheme in UniDic consists of roughly 2 to 3 characters. Therefore, six words and an omitted morpheme roughly correspond to 20 characters.

Note that the target abbreviations and the correct full forms were excluded from the feature.

2.2 Materials

Comparison of these two methods required the preparation of two materials: 1) an ambiguous abbreviation set and 2) a training set.

[a] MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourcrforge.net/. Accessed on Oct. 26, 2011

[b] Unidic. http://www.tokuteicorpus.jp/dist/. Accessed on Oct. 26, 2012

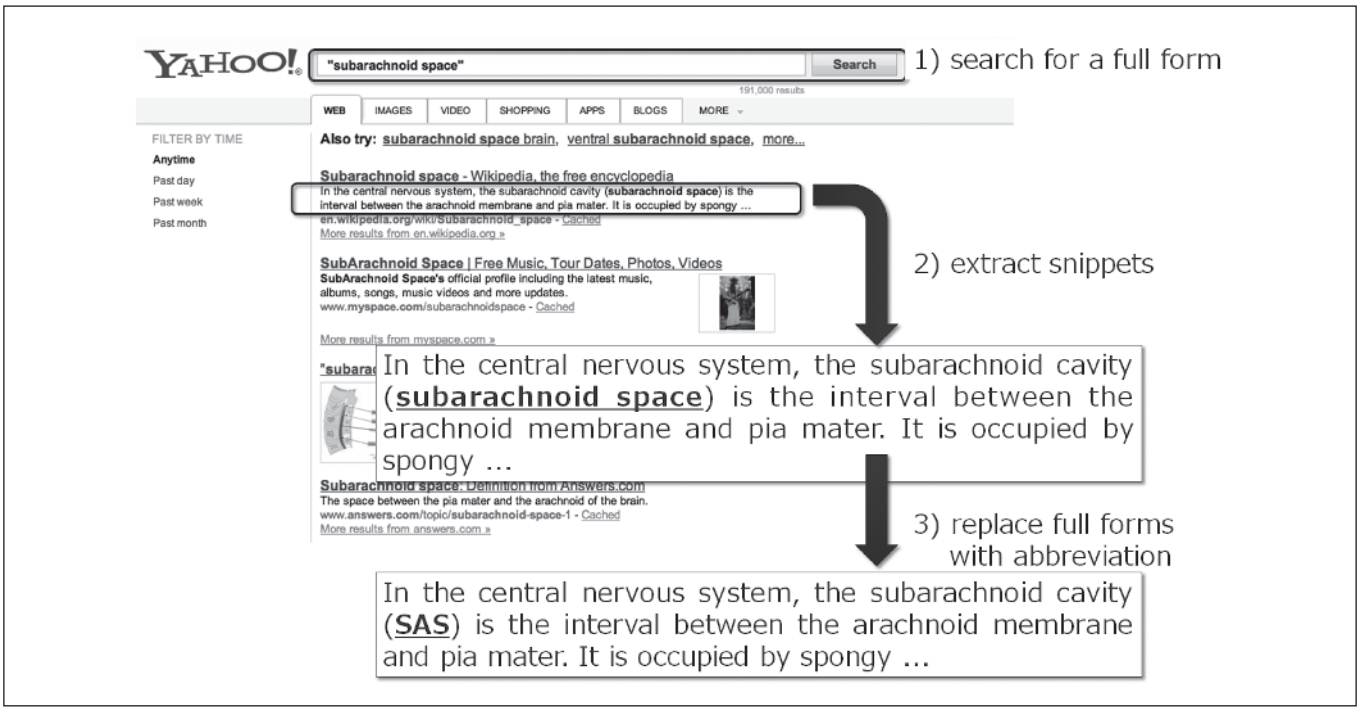

Figure 6 Construction of a pseudo-annotated corpus. Construction entailed the following three steps: 1) search for a full form on the Internet (the actual construction was done through API provided by Yahoo! Japan); 2) extract snippets from the search results; and 3) replace all queries (full form) that appeared in the snippets with their abbreviations.

2.2.1 Ambiguous Abbreviation Set (<abbreviation, full forms> Pairs)

We chose an ambiguous abbreviation set as an experimental target. A single abbreviation would be insufficient for comparison purposes because accuracies are expected to differ depending on the abbreviations used. Moreover, the target abbreviations are required to have specific characteristics. First, they must have more than one full form. Second, a certain quantity of experimental data is necessary for each full form. We arbitrarily set the threshold to 10. Thus, we set the selection criteria as follows:

1. Abbreviations must correspond to two or more full forms in the medical abbreviation dictionary [34].
2. All corresponding Japanese full forms must have appeared in at least 10 web pages.
3. All corresponding Japanese full forms had to be at least 5 characters in length.

The third criterion was used due to the reduced ambiguity required for construction of a corpus with high quality, as described in the next section (2.2.2 Corpus).

2.2.2 Corpus

The implementation of the two methods and measurement of their performance required an annotated corpus. An annotated corpus is a set of texts in which abbreviations appear and for which their correct full forms are annotated. Building such a corpus by hand, however, is extremely costly and time-consuming. Instead, we constructed a pseudo-annotated corpus assuming that an abbreviation and its full form are used in the same context [23]. As shown in ▶Figure 6, the construction of the corpus required three steps:

1. Search for full forms on the Internet
2. Extract snippets
3. Replace full forms in the snippets with corresponding abbreviations

We used Yahoo! API as a Web search engine to search the Internet. The API received search queries and returned the following results: the total number of hits; a set of URLs with the titles; and the abstracts (called "snippets"). Finally, we consolidated all the collected snippets and eliminated any duplicates.

For example, a corpus of SAS as an abbreviation of "subarachnoid space" was constructed in the following manner. First, we searched for "subarachnoid space" on the Internet via Yahoo! API. Next, we extracted snippets from the result. Finally, we replaced "subarachnoid space" appearing in each snippet with the abbreviation SAS. The resulting sentences were samples of documents in which SAS appeared as an abbreviation of "subarachnoid space". The actual query for the full form was in Japanese, not English.

While we did not explicitly restrict the search domain, the queries were thought to be adequately specific in medical content which implicitly restricted the search.

2.3 Evaluation Scale

We used expansion accuracy as the evaluation scale. Accuracy was defined as the rate of correctly expanded samples among all samples. Expansion accuracy was calculated for each target abbreviation. Evaluation was conducted in a five-fold cross-validation manner. The accuracy could be compatible with precision because the sys-

Table 1 Experimental data. Eight ambiguous abbreviations selected based on the selection criterion, each corresponding to five to seven full forms. Abbr., #abbr. and #data indicate abbreviation, the number of the abbreviations in the set of snippets (after the corpus construction step 2), and the number of abbreviations in the final corpus, respectively. #data is generally much larger than #abbr because of the substitution for the full spellings for the abbreviations.

abbr.	full form	#abbr.	#data
ASA	acetylsalicylic acid	3	398
	active systemic anaphylaxis	5	18
	anti-smooth muscle antibody	1	346
	argininosuccinic acid	2	320
	aspirin sensitive asthma	4	560
DHA	dehydro ascorbic acid	41	371
	dehydroacetic acid	4	452
	dehydroepiandrosterone	2	327
	dihydroxyadenine	8	31
	docosahexaenoic acid	353	688
DIC	adipiodone meglumine	0	8
	disseminated intravascular coagulation	333	719
	drip infusion cholangiography	37	69
	drip infusion cholecystocholangiography	13	27
	drip infusion cholecystograph	26	57
PAN	periarteritis nodosa	6	408
	periodic alternating nystagmus	1	10
	polyacrylonitrile	131	491
	polyarteritis nodosa	6	302
	puromycin aminonucleoside nephrosis	2	21
PCI	percutaneous coronary intervention	296	628
	peripheral circulatory impairment	0	462
	pneumatosis cystoides intestinalis	1	97
	prophylactic cranial irradiation	91	375
	protein C inhibitor	18	55
PID	pelvic inflammatory disease	150	150
	phenindione	0	27
	plasma iron disappearanc	31	85
	primary immunodeficiency	30	427
	prolapsed intervertebral disk	0	655
PPP	palatopharyngoplasty	25	80
	pancreatic polypeptide	1	111
	pentose phosphate pathway	0	65
	pigmented pretibial patches	2	29
	platelet poor plasma	64	179
	platelet poor plasma	104	507
	pustulosis palmaris et plantaris	4	451
SAS	aortic stenosis subaortic stenosis	0	277
	sleep apnea syndrome	163	730
	subarachnoid space	0	526
	supravalvular aortic stenosis	1	40
	sympathicoadrenal system	0	5

Table 2 Accuracy of abbreviation expansion obtained using two methods* indicates that ***method C*** performed significantly better than ***method M*** ($p < 0.05$).

	ASA	DHA	DIC	PAN*	PCI	PID*	PPP*	SAS
method M	**0.892**	0.910	0.932	0.796	0.895	0.894	0.893	0.923
method C	0.890	**0.912**	**0.942**	**0.836**	**0.906**	**0.931**	**0.914**	**0.938**

tem outputs exact one full form for all the input abbreviations.

3. Results

Based on the selection criteria, eight abbreviations and 42 full forms were identified. The constructed corpus contained 276 samples on average per full form and the total size was 11584. Details are provided in ▶Table 1. ▶Table 2 presents the results from the evaluation. Accuracies for ***method C*** exceeded ***method M*** for all abbreviations except ASA. McNemar's test indicated that accuracies for ***method C*** were significantly higher for three out of the eight abbreviations ($p < 0.05$). Accuracy information for the remaining five abbreviations did not yield significant differences between the two methods.

▶Figure 7 shows associations between window sizes and expansion accuracies for ***method C***. Although accuracies depend on abbreviations, they share a tendency to increase along with window size.

4. Discussion

The present results revealed that omitting morphological analysis improved the performance of abbreviation expansion in some cases. *Method C* was superior to *method M* for three out of the eight abbreviations. While it is difficult to claim the

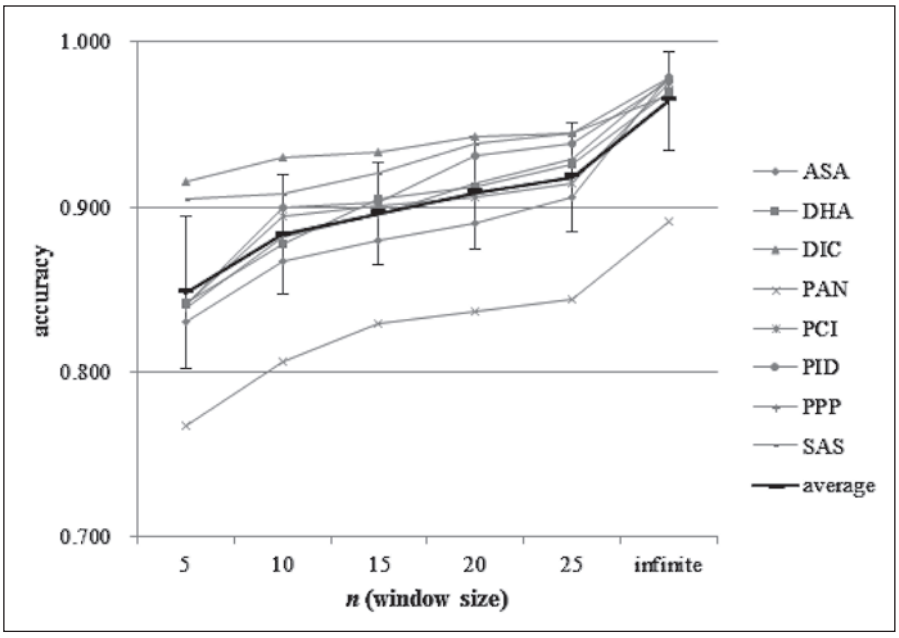

Figure 7 Relationship between window size and accuracy. Accuracy increases along with window size. The error range for the average is the standard deviation. "Infinite" means "entire snippet text."

superiority of *method C* in general, the present results suggest that *method C* was not worse than *method M*. In terms of its simplicity of implementation, *method C* is superior to *method M* because it requires no additional processing such as morphological analysis except to simply split a sentence into its characters. Thus, all things considered, our results suggest that *method C* is superior to *method M*.

Note that the present study compared character information with available morphological information, not ideal one. As described in Introduction, ideal morphological information of Japanese text is not available at the present day. This means that no one knows the ideal morpheme sequences, let alone a morphological analyzer.

As compared to *method C*, *method M* has two challenges -- quality of the dictionary and the performance of the morphological analyzer. One would guess the latter probably decreased the performance of *method M*, but the above situation makes construction of a morphological annotated corpus difficult, which is required by the evaluation of the performance.

The clear mistakes of the morphological analyzer in the experiment were caused by dirtied snippets: snippets including unwanted characters (e.g. "respo.nse") and snippets whose ending morpheme was cut (e.g. "in the absence of resp…"). For the dirtied input, *method C* is clearly more robust than *method M*.

In order to further examine the factors which led to *method C* working better for the three abbreviations, we investigated the feature characteristics of the cases for which only *method M* failed. First, we separated the data into two groups (data-centered) and counted each feature for each case (e.g. the number of noun appeared preceding abbreviation that used as a part of input feature to SVM): dC, cases that only *method C* outputted correct answers, and dO, the other cases (▶Table 3). In dC, surrounding nouns are significantly fewer than that in dO (paired t-test, $p = 0.01$). That is, the lack of nouns in the context would be the reason that only *method M* failed. Next, we separated the data in a different manner (abbreviation-centered) and did frequency counts of zero for each feature (e.g. the number of cases without nouns preceding the abbreviation): aC, for all cases of the three abbreviations for which *method C* worked better, and aO the other data (▶Table 4). In aC, there were significantly fewer cases without preceding than in aO. Therefore, the reason why *method C* is superior to *method M* for only three abbreviations is thought to be the high ratio of the lack of a left noun context.

In general, morphological analysis plays an important role in NLP. Many applications, such as dependency parsing, named entity recognition and information retrieval, are often based on a given morpheme sequence. In some cases, however, it is not essential. Abbreviation expansion is one such case, as shown in this paper. An additional example may be to search within documents for technical terms in dictionaries because simple string matching to

Table 3 Comparison of feature frequency (data-centered). This table shows the average frequency for each feature with respect to their abbreviations andgroups dC and dO. V1, V2, N1, and N2 indicates a verb preceding an abbreviation, a verb following an abbreviation, a noun preceding an abbreviation, and a noun following an abbreviation, respectively. Paired t-tests showed significant difference on N1 and N2 between dC and dO.

	dC				dO			
abbr.	V1	V2	N1	N2	V1	V2	N1	N2
ASA	0.9	2.1	2.2	2.9	0.9	1.8	2.5	2.9
DHA	0.9	1.5	2.4	2.9	0.7	1.7	2.4	2.9
DIC	0.4	1.2	2.0	2.8	0.7	1.8	2.3	2.9
PAN	0.5	1.3	2.0	2.9	0.6	1.2	2.3	2.9
PCI	0.7	1.6	2.3	2.8	0.9	1.8	2.6	2.9
PID	0.9	1.8	2.3	2.8	0.8	1.9	2.3	2.9
PPP	0.7	2.0	2.2	2.9	1.0	2.2	2.3	2.9
SAS	1.2	2.0	2.6	2.9	1.1	2.2	2.4	2.9

words in dictionaries is often sufficient. Technical terms often are long and characteristic in their spelling, which makes the word unambiguous. However, such a limited purpose does not require the power of morphological analysis.

While accuracies for abbreviation were high, accuracies for full forms varied. Accuracy was low 1) when data were relatively scarce and 2) when there were similar full forms. The first reason is not surprising, as the quantity of training data is generally correlated with performance. In contrast, the second reason was unexpected. We expected that full forms corresponding to the same abbreviation would not be similar to each other; however, the results from the present study showed that this was not the case. The implication was that one should carefully construct a full form inventory; for example, simplify the inventory by aggregating similar full forms and, if necessary, preparing additional models to distinguish among them.

Because *method C* does not require information from words including their POS which is required by *method M*, it is applicable to all languages. As we set the character features for Japanese, it is necessary to adjust the parameters in order to apply this method to other languages. Roughly estimating, the size of the feature space was approximately 7,000 because uni-/bi-gram of Hiragana and Katakana are 5,000 as they each have 70 characters, and there are about 2000 Kanji. English has 26 characters, thus, in order to construct a feature space size of 7,000, a bi-/tri-gram should be used. However, the results from the present study do not guarantee the performance of *method C* for all languages.

Table 4 Comparison of zero frequency of features (abbreviation-centered). This table shows the average frequency for cases without each feature with respect to the abbreviations grouped by aC and aO. V1, V2, N1, and N2 are the same meaning in Table 3. T-tests showed the significant difference on N1 between aC and aO.

	abbr.	V1	V2	N1	N2
aC	PAN	0.72	0.49	0.20	0.01
	PID	0.60	0.21	0.17	0.01
	PPP	0.55	0.16	0.18	0.01
aO	ASA	0.58	0.25	0.14	0.01
	DHA	0.64	0.32	0.14	0.01
	DIC	0.64	0.27	0.14	0.01
	PCI	0.57	0.26	0.09	0.01
	SAS	0.50	0.12	0.15	0.01

4.1 Limitation: Characteristics of the Corpus

The corpus has two characteristics compared in clinical text: 1) the domain and 2) the length of each datum.

Although our actual target was the health care domain, the experimental data were collected from the Internet. The diversity of the snippets was not clear. While some describe definitions of the full form, others described a famous person suffering from disease or individual experiences. While judging each occurrence technicality would not be easy, the number of hits may serve as an indicator, which is given explicitly by the search engine. Expectedly, the more hits, the more popular the full form was – that is, the lower the technicality. Our corpus included A) four full forms that with more than 10,000 hits; B) 12 full forms with 1,000 to 9,999 hits; and C) 26 full forms with less than 1,000 hits. The majority of the occurrences found were news and individual experiences in group A, definitions and explanations in group B and scientific documents in group C. Whatever the distribution of snippets may be, the content of the corpus is more or less different from that of clinical documents. Therefore, these results cannot be directly applied to clinical documents. Further investigation is required. Moreover, our corpus is more likely comprised of local abbreviations that are explicitly defined in the document by nature, while our target is global abbreviation. This point is further discussed in the section Limitation: Strong assumption.

Because the corpus was automatically constructed using a Web search engine, the length of each datum was fully dependent on how the search engine constructed snippets. Still, we did not consider this as a fatal limitation of the experiment. Table 5 presents the average length of the context for each abbreviation in our corpus. The left verb/noun contexts were less than three, while ***method M*** relied on three verbs and nouns on the left and right. This could be a disadvantage for ***method M***; however, there was no assurance that there would be enough context for the actual situation. This is an intrinsic limitation of morpheme-based methods rather than the

Table 5 Context length of the corpus. This table displays the number of morphemes/characters to the abbreviation's left/right. Note that ***method M*** uses three verbs and nouns and ***method C*** uses n characters for context, from left/right, respectively.

	#morphemes				#characters	
	left (verb)	left (noun)	right (verb)	right (noun)	left	right
ASA	1.0	8.9	2.2	14.8	39.8	78.7
DHA	0.8	7.7	1.6	13.5	40.6	79.0
DIC	0.5	7.5	1.6	17.1	35.5	87.6
PAN	0.7	8.8	1.4	20.2	41.1	91.3
PCI	1.1	9.6	1.8	17.5	40.1	81.5
PID	1.1	8.8	1.8	16.4	38.0	81.7
PPP	1.2	10.1	2.9	17.2	39.7	83.4
SAS	1.0	9.6	3.3	16.4	35.1	78.8

present experiment. On the other hand, character context on both the left and right are more than 20 characters in length on the same corpus, which corresponded to approximately 90% accuracy.

4.2 Limitation: Strong Assumption

We adopted the assumption that an abbreviation and its full form are used in the same context, based on which we constructed a pseudo-annotated corpus. However, this assumption is not always correct, especially for professional documents such as clinical notes; documents that contain abbreviations may include abbreviations of many kinds. Therefore, it is non-trivial that the result can be applicable to such documents. Experiments using clinical text are necessary; however, this was outside the scope of the present study, since we did not have enough clinical text that contained ambiguous abbreviations or full forms.

The pseudo-annotated corpus based on this assumption has one more limitation: it requires an additional assumption that the query is not ambiguous. Some full forms, such as "acid", are actually ambiguous, while other full forms have little currency. In such circumstances, one should build an annotated corpus by hand for all possible abbreviations. This is not a realistic approach; the supervised approach is unsuitable for construction of practicable applications. Instead, semi-/un-supervised methods such as a dictionary-based method [32] might be appropriate. For such an approach, the results from the present study are expected to contribute to the existing literature.

5. Conclusions

The present preliminary study investigated whether morphological analysis influenced the performance of abbreviation expansion by comparing a morpheme-based method (*method M*) to a character-based method (*method C*) using English abbreviations appearing in Japanese text extracted from documents on the World Wide Web. The results demonstrated that *method C* yielded significantly better results than *method M* for three out of the eight abbreviations. No significant differences were found for the remaining five abbreviations. Considering the simplicity of implementation, we concluded that *method C* is superior to *method M*.

Morphological analysis is often a source of problems in Japanese-language texts, especially within the health care domain. We hope that the results from the present study help to resolve difficulties caused by analysis errors and morpheme definitions.

Acknowledgment

The present study was undertaken as part of a joint research effort with the University of Tokyo Hospital, Fuji Xerox. The University of Tokyo Hospital has received research funding from Fuji Xerox, where EY Shinohara is employed.

References

1. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. AMIA Summits Transl Sci Proc. 2010; 2010: 1–5.
2. Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? J Biomed Inform 2009; 42 (5): 760–772.
3. Aramaki E, Miura Y, Tonoike M, Ohkuma T, Masuichi H, Waki K, et al. Extraction of adverse drug effects from clinical records. Stud Health Technol Inform 2010; 160 (Pt 1): 739–743.
4. Stetson PD, Johnson SB, Scotch M, Hripcsak G. The sublanguage of cross-coverage. Proc AMIA Symp; 2002. pp 742–746.
5. Chase HS, Kaufman DR, Johnson SB, Mendonca EA. Voice capture of medical residents' clinical information needs during an inpatient rotation. J Am Med Inform Assoc 2009; 16 (3): 387–394.
6. Minard AL, Ligozat AL, Ben Abacha A, Bernhard D, Cartoni B, Deléger L, et al. Hybrid methods for improving information access in clinical documents: concept, assertion, and relation identification. J Am Med Inform Assoc. 2011; 18 (5): 588–593.
7. Compliance data for the Joint Commissions' 2004 and 2005 national patient safety goals. Jt Comm Perspect 2005; 25: 7–8.
8. Myers JS, Gojraty S, Yang W, Linsky A, Airan-Javia S, Polomano RC. A randomized-controlled trial of computerized alerts to reduce unapproved medication abbreviation use. J Am Med Inform Assoc 2011; 18 (1): 17–23.
9. Gaudan S, Kirsch H, Rebholz-Schuhmann D. Resolving abbreviations to their senses in Medline. Bioinformatics 2005; 21 (18): 3658–3664.
10. Yu H, Hripcsak G, Friedman C. Mapping abbreviations to full forms in biomedical articles. J Am Med Inform Assoc 2002; 9 (3): 262–272.
11. Xu H, Stetson PD, Friedman C. A study of abbreviations in clinical notes. AMIA Annu Symp Proc; 2007. pp 821–825.
12. Xu H, Stetson PD, Friedman C. Methods for building sense inventories of abbreviations in clinical notes. J Am Med Inform Assoc 2009; 16 (1): 103–108.
13. Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pac Symp Biocomput; 2003. pp 451–462.
14. Okazaki N, Ananiadou S. Building an abbreviation dictionary using a term recognition approach. Bioinformatics 2006; 22 (24): 3089–3095.
15. Wren JD, Garner HR. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. Methods Inf Med. 2002; 41 (5): 426–434.
16. Ao H, Takagi T. ALICE: an algorithm to extract abbreviations from MEDLINE. J Am Med Inform Assoc 2005; 12 (5): 576–586.
17. Gale WA, Church KW, Yarowsky D. A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities 1992; 26: 415–439.
18. Brown PF, Pietra SAD, Pietra VJD, Mercer RL. Word-sense disambiguation using statistical methods. Proceedings of the 29th annual meeting on Association for Computational Linguistics; Berkeley, California: Association for Computational Linguistics; 1991.
19. Lesk M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. Proceedings of the 5th annual international conference on Systems documentation; Toronto, Ontario, Canada: ACM; 1986.
20. Guthrie JA, Guthrie L, Wilks Y, Aidinejad H. Subject-dependent co-occurrence and word sense disambiguation. Proceedings of the 29th annual meeting on Association for Computational Linguistics; Berkeley, California: Association for Computational Linguistics; 1991.
21. Yarowsky D. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. Proceedings of the 14th conference on Computational linguistics – Volume 2; Nantes, France: Association for Computational Linguistics; 1992.
22. Dagan I, Itai A. Word sense disambiguation using a second language monolingual corpus. Comput Linguist 1994; 20 (4): 563–596.
23. Pakhomov S. Semi-supervised Maximum Entropy based approach to acronym and abbreviation normalization in medical texts. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics; Philadelphia, Pennsylvania: Association for Computational Linguistics; 2002.
24. Pakhomov S, Pedersen T, Chute CG. Abbreviation and acronym disambiguation in clinical discourse. AMIA Annu Symp Proc; 2005. pp 589–593.
25. Joshi M, Pakhomov S, Pedersen T, Chute CG. A comparative study of supervised learning as applied to acronym expansion in clinical reports. AMIA Annu Symp Proc; 2006. pp 399–403.
26. Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. J Am Med Inform Assoc 2004; 11 (4): 320–331.

27. Savova GK, Coden AR, Sominsky IL, Johnson R, Ogren PV, de Groen PC, et al. Word sense disambiguation across two domains: biomedical literature and clinical notes. J Biomed Inform 2008; 41 (6): 1088–1100.
28. Liu H, Johnson SB, Friedman C. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. J Am Med Inform Assoc 2002; 9 (6): 621–636.
29. Yamada E, Aramaki E, Tonoike M, Ohkuma T, Miura Y, Sugihara D, Masuichi H, and Ohe K. Abbreviation Disambiguation in Japanese Medical Text. Jpn J Med Inf 2010; 30 (Suppl.): 389–392. In Japanese.
30. Schutze H. Word sense disambiguation with sublexical representations. Proc. Workshop on Statistically-Based NLP Techniques, AAAI Technical Report WS-92–01; 1992. pp 100–104.
31. Schutze H. Automatic word sense discrimination. Comput Linguist 1998; 24 (1): 97–123.
32. Yarowsky D. Unsupervised word sense disambiguation rivaling supervised methods. Proceedings of the 33rd annual meeting on Association for Computational Linguistics; Cambridge, Massachusetts: Association for Computational Linguistics; 1995.
33. Okazaki N, Ananiadou S, and Tsujii J. Building a High Quality Sense Inventory for Improved Abbreviation Disambiguation. Bioinformatics 2010; 26:9: 1246–1253.
34. Japan Collegium on Hospital Administration. 16000 Abbreviations in Medical record & Receipt. Igakutushinsya Co. Ltd.; 2008. ISBN 978-4-87058-367-2. In Japanese.
35. Nishimoto N, Terae S, Uesugi M, Ogasawara K, Sakurai T. Development of a medical-text parsing algorithm based on character adjacent probability distribution for Japanese radiology reports. Methods Inf Med 2008; 47 (6): 513–521.
36. Baldwin T, Kim SN, Bond F, Fujita S, Martinez D, Tanaka T. MRD-based word sense disambiguation: further extending LESK. In: Proceedings of the 3rd International Joint Conference on Natural Language Processing; 2008. pp 775–780.
37. Fujii H, Croft WB. A comparison of indexing techniques for Japanese text retrieval. Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval; Pittsburgh, Pennsylvania, United States: ACM; 1993.
38. Baldwin T. Low-cost, high-performance translation retrieval: dumber is better. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics; Toulouse, France: Association for Computational Linguistics; 2001.
39. MeCab: Yet Another Part-of-Speech and Morphological Analyzer. http://mecab.sourceforge.net/ Accessed Oct. 26, 2011.
40. Unidic. http://www.tokuteicorpus.jp/dist/ Accessed Oct. 26, 2011.