A Method of Measuring Term Representativeness
- Baseline Method Using Co-occurrence Distribution -

Toru Hisamitsu,† Yoshiki Niwa,† and Jun-ichi Tsujii‡

† Central Research Laboratory, Hitachi, Ltd.
Akanuma 2520, Hatoyama, Saitama 350-0395, Japan
{hisamitu, yniwa}@harl.hitachi.co.jp

‡ Graduate School of Science, the University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan
tsujii@is.s.u-tokyo.ac.jp

Abstract

This paper introduces a scheme, which we call the baseline method, to define a measure of term representativeness and measures defined by using the scheme. The representativeness of a term is measured by a normalized characteristic value defined for a set of all documents that contain the term. Normalization is done by comparing the original characteristic value with the characteristic value defined for a randomly chosen document set of the same size. The latter value is estimated by a baseline function obtained by random sampling and logarithmic linear approximation. We found that the distance between the word distribution in a document set and the word distribution in a whole corpus is an effective characteristic value to use for the baseline method. Measures defined by the baseline method have several advantages including that they can be used to compare the representativeness of two terms with very different frequencies, and that they have well-defined threshold values of being representative. In addition, the baseline function for a corpus is robust against differences in corpora; that is, it can be used for normalization in a different corpus that has a different size or is in a different domain.

1 Introduction

Measuring the representativeness (i.e., the informativeness or domain specificity) of a term[1] is essential to various tasks in natural language processing (NLP) and information retrieval (IR). It is particularly crucial when applied to an IR interface to help a user find informative terms. For instance, when the number of retrieved documents is intractably large, an overview of representative words in the documents is needed to understand the contents. To enable this, an IR system, called *DualNAVI,* that has two navigation windows where one displays a graph of representative words in the retrieved documents, was developed (Nishioka et al. 1997). This window helps users grasp the contents of retrieved documents, but it also exposes problems concerning existing representativeness measures.

Figure 1 shows an example of a graph for the query 電子マネー (electronic money), with *Nihon Keizai Shimbun* (a financial newspaper) 1996 as the corpus. Frequently appearing words are displayed in the upper part of the window, and words are selected by a *tf-idf*-like measure (Niwa et al. 1997). Typical non-representative words are filtered out by using a stop-word list.

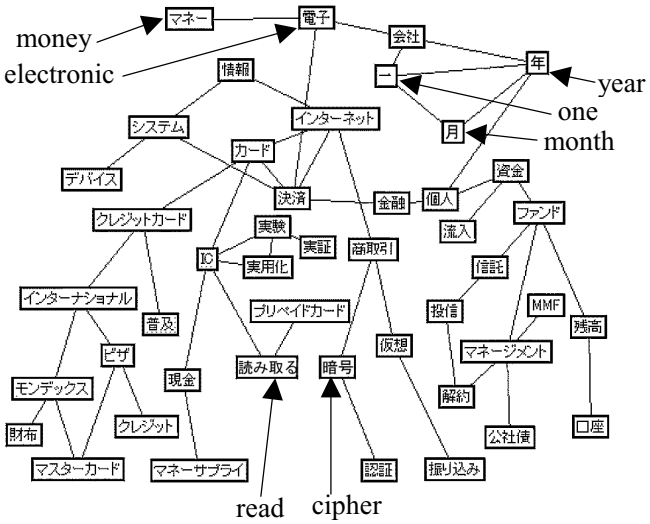

Figure 1
A topic word graph when the query is 電子マネー(electronic money).

One problem is the difficulty of suppressing uninformative words such as 年 (year), 一 (one), and 月 (month) because classical measures, such as *tf-idf*, are too sensitive to word frequency and no established method to automatically construct a stop-word list has been developed.

Another problem is that the difference in the representativeness of words is not sufficiently indicated. In the example above, highlighting 暗号 (cipher) over less representative words such as 読みとる (read) would be useful. Most classical measures based on only term frequency and document frequency cannot overcome this problem.

To define a more elaborate measure, attempts to incorporate more precise co-occurrence information have been made. Caraballo et al. (1999) tried to define a measure for "specificity" of a noun by using co-occurrence information of a noun, but it was not very successful in the sense that the measure did not particularly outperformed the term frequency.

Hisamitsu et al. (1999) developed a measure of the representativeness of a term by using co-occurrence information and a normalization

[1] A term is a word or a word sequence.

technique. The measure is based on the distance between the word distribution in the documents containing a term and the word distribution in the whole corpus. Their measure overcomes previously mentioned problems and preliminary experiments showed that this measure worked better than existing measures in picking out representative/non-representative terms. Since the normalization technique plays a crucial part of constructing the measure, issues related to the normalization need more study.

In this paper we review Hisamitsu's measure and introduce a generic scheme -- which we call the baseline method for convenience -- that can be used to define various measures including the above. A characteristic value of all documents containing a term T is normalized by using a baseline function that estimates the characteristic value of a randomly chosen document set of the same size. The normalized value is then used to measure the representativeness of the term T. A measure defined by the baseline-method has several advantages compared to classical measures.

We compare four measures (two classical ones and two newly defined ones) from various viewpoints, and show the superiority of the measure based on the normalized distance between two word distributions. Another important finding is that the baseline function is substantially portable, that is, one defined for a corpus can be used for a different corpus even if the two corpora have considerably different sizes or are in different domains.

2. Existing measures of representativeness

2.1 Overview

Various methods for measuring the informativeness or domain specificity of a word have been proposed in the domains of IR and term extraction in NLP (see the survey paper by Kageura 1996). In characterizing a term, Kageura introduced the concepts of "unithood" and "termhood": unithood is "the degree of strength or stability of syntagmatic combinations or collocations," and termhood is "the degree to which a linguistic unit is related to (or more straightforwardly, represents) domain-specific concepts." Kageura's termhood is therefore what we call representativeness here.

Representativeness measures were first introduced in an IR domain for determining indexing words. The simplest measure is calculated from only word frequency within a document, For example, the weight I_{ij} of word w_i in document d_j is defined by

$$I_{ij} = \frac{f_{ij}}{\sum_k f_{kj}},$$

where f_{ij} is the frequency of word w_i in document d_j (Sparck-Jones 1973, Noreault et al. 1977). More elaborate measures for termhood combine word frequency within a document and word occurrence over a whole corpus. For instance, *tf-idf*, the most commonly used measure, was originally defined as

$$I_{ij} = f_{ij} \times \log(\frac{N_{total}}{N_i}),$$

where N_i and N_{total} are, respectively, the number of documents containing word w_i and the total number of documents (Salton et al. 1973). There are a variety of definitions of *tf-idf*, but its basic feature is that a word appearing more frequently in fewer documents is assigned a higher value. If documents are categorized beforehand, we can use a more sophisticated measure based on the χ^2 test of the hypothesis that an occurrence of the target word is independent of categories (Nagao et al. 1976).

Research on automatic term extraction in NLP domains has led to several measures for weighting terms mainly by considering the unithood of a word sequence. For instance, mutual information (Church et al. 1990) and the log-likelihood (Dunning 1993) methods for extracting word bigrams have been widely used. Other measures for calculating the unithood of n-grams have also been proposed (Frantzi et al. 1996, Nakagawa et al. 1998, Kita et al. 1994).

2.2 Problems

Existing measures suffer from at least one of the following problems:

(1) Classical measures such as *tf-idf* are so sensitive to term frequencies that they fail to avoid very frequent non-informative words.
(2) Methods using cross-category word distributions (such as the χ^2 method) can be applied only if documents in a corpus are categorized.
(3) Most measures in NLP domains cannot treat single word terms because they use the unithood strength of multiple words.
(4) The threshold value for being representative is defined in an *ad hoc* manner.

The scheme that we describe here constructs measures that are free of these problems.

3. Baseline method for defining representativeness measures

3.1 Basic idea

This subsection describes the method we developed for defining a measure of term representativeness. Our basic idea is summarized by the famous quote (Firth 1957) :

> "*You shall know a word by the company it keeps*."

We interpreted this as the following working hypothesis:

> For any term T, if the term is representative, $D(T)$, the set of all documents containing T, should have some characteristic property compared to the "average".

To apply this hypothesis, we need to specify a measure to obtain some "property" of a document set and the concept of "average". Thus, we converted this hypothesis into the following procedure:

> Choose a measure M characterizing a document set. For term T, calculate $M(D(T))$, the value of the measure for $D(T)$. Then compare $M(D(T))$ with $B_M(\#D(T))$, where $\#D(T)$ is the number of words contained in $\#D(T)$, and B_M estimates the value of $M(D)$ when D is a randomly chosen document set of size $\#D(T)$.

Here, M measures the property and B_M estimates the average. The size of a document set is defined as the number of words it contains.

We tried two measures as M. One was the number of different words (referred to here as *DIFFNUM*) appearing in a document set. Teramoto conducted an experiment with a small corpus and reported that *DIFFNUM* was useful for picking out important words (Teramoto et al. 1999)[2] under the hypothesis that the number of different words co-occurring with a topical (representative) word is smaller than that with a generic word. The other measure was the distance between the word distribution in $D(T)$ and the word distribution in the whole corpus D_0. The distance between the two distributions can be measured in various ways, and we used the log-likelihood ratio as in Hisamitsu et al. 1999, and denote this measure as *LLR*. Figure 2 plots $(\#D, M(D))$s when M is *DIFFNUM* or *LLR*, where D varies over sets of randomly selected documents of various sizes from the articles in *Nikkei-Shinbun* 1996.

For measure M, we define $Rep(T, M)$, the representativeness of T, by normalizing $M(D(T))$ by $B_M(\#D(T))$. The next subsection describes the construction of B_M and the normalization.

3.2 Baseline function and normalization

Using the case of *LLR* as an example, this subsection explains why normalization is necessary and describes the construction of a baseline function.

Figure 3 superimposes coordinates $\{(\#D(T), LLR(D(T))\}$s onto the graph of *LLR* where T varies

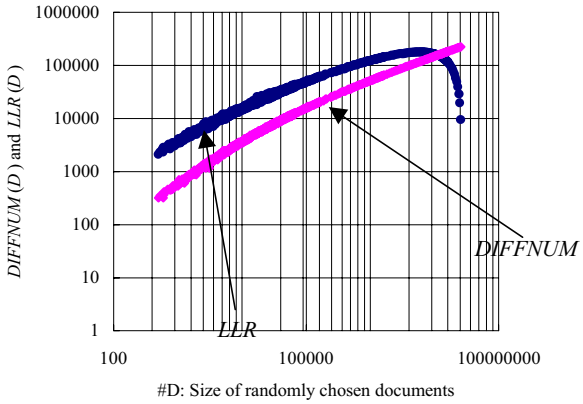

Figure 2
Values of *DIFFNUM* and *LLR* for randomly chosen document set.

over 暗号(cipher), 年(year), 月(month), 読み取る(read), 一(one), する(do), and 経済(economy). Figure 3 shows that, for example, $LLR(D(\text{する}))$ is smaller than $LLR(D(\text{経済}))$, which reflects our linguistic intuition that words co-occurring with "economy" are more biased than those with "do". However, $LLR(D(\text{暗号}))$ is smaller than $LLR(D(\text{読み取る}))$ and smaller even than $LLR(D(\text{する}))$. This contradicts our linguistic intuition, and is why values of *LLR* are not directly used to compare the representativeness of terms. This phenomenon arises because $LLR(D(T))$ generally increases as $\#D(T)$ increases. We therefore need to use some form of normalization to offset this underlying tendency.

We used a baseline function to normalize the values. In this case, $B_{LLR}(\bullet)$ was designed so that it approximates the curve in Fig. 3. From the definition of the distance, it is obvious that $B_{LLR}(0) = B_{LLR}(\#D_0) = 0$. At the limit when $\#D_0 \to \infty$, $B_{LLR}(\bullet)$ becomes a monotonously increasing function.

The curve could be approximated precisely through logarithmic linear approximation near (0, 0). To make an approximation, up to 300 documents are randomly sampled at a time. (Let each randomly chosen document set be denoted by D. The number of sampled documents are increased from one to 300, repeating each number up to five times.) Each $(\#D, LLR(D))$ is converted to $(\log(\#D), \log(LLR(D)))$. The curve formulated by the $(\log(\#D), \log(LLR(D)))$ values, which is very close to a straight line, is further divided into multiple parts and is part-wise approximated by a linear function. For instance, in the interval $I = \{x \mid 10000 \le x < 15{,}000\}$, $\log(LLR(D))$ could be approximated by $1.103 + 1.023 \times \log(\#D)$ with $R^2 = 0.996$.

For *LLR*, we define $Rep(T, LLR)$, the representativeness of T by normalizing $LLR(D(T))$ by $B_{LLR}(\#D(T))$ as follows:

$$Rep(T, LLR) = 100 \times \left(\frac{\log(LLR(D(T)))}{\log(B_{LLR}(\#D(T)))} - 1\right).$$

[2] With Teramoto's method, eight parameters must be tuned to normalize $DIFFNUM(D(T))$, but the details of how this was done were not disclosed.

For instance, when we used *Nihon Keizai Shimbun* 1996, The average of 100×(log(*LLR*(*D*)) /log(B_{LLR} (#*D*)) − 1), *Avr*, was −0.00423 and the standard deviation, σ, was about 0.465 when *D* varies over randomly selected document sets. Every observed value fell within *Avr*±4σ and 99% of observed values fell within Avr±3σ. This happened in all corpora (7 orpora) we tested. Therefore, we can define the threshold of being representative as, say, *Avr* + 4σ.

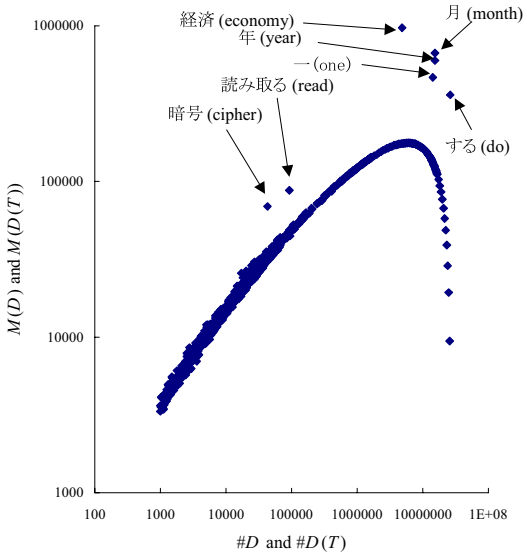

Figure 3
Baseline and sample word distribution

3.3 Treatment of very frequent terms

So far we have been unable to treat extremely frequent terms, such as する (do). We therefore used random sampling to calculate the *Rep*(*T, LLR*) of a very frequent term *T*. If the number of documents in *D*(*T*) is larger than a threshold value *N*, which was calculated from the average number of words contained in a document, *N* documents are randomly chosen from *D*(*T*) (we used *N* = 150). This subset is denoted $\underline{D}(T)$ and *Rep*(*T, LLR*) is defined by 100 × (log(*LLR*($\underline{D}(T)$)) /log(B_{LLR} (#$\underline{D}(T)$)) − 1). This is effective because we can use a well-approximated part of the baseline curve; it also reduces the amount of calculation required.

By using *Rep*(*T, LLR*) defined above, we obtained *Rep*(する, *LLR*) = −0.573, *Rep*(読み取る, *LLR*) = 4.08, and *Rep*(暗号, *LLR*) = 6.80, which reflect our linguistic intuition.

3.4 Features of *Rep*(*T*, *M*)

Rep(*T*, *M*) has the following advantages by virtue of its definition:

(1) Its definition is mathematically clear.
(2) It can compare high-frequency terms with low-frequency terms.
(3) The threshold value of being representative can be defined systematically.
(4) It can be applied to *n*-gram terms for any *n*.

4. Experiments

4.1 Evaluation of monograms

Taking topic-word selection for a navigation window for IR (see Fig. 1) into account, we examined the relation between the value of *Rep*(*T*, *M*) and a manual classification of words (monograms) extracted from 158,000 articles (excluding special-styled non-sentential articles such as company-personnel-affair articles) in the 1996 issues of the *Nikkei Shinbun*.

4.1.1 Preparation

We randomly chose 20,000 words from 86,000 words having document frequencies larger than 2, then randomly chose 2,000 of them and classified these into three groups: *class a* (acceptable) words useful for the navigation window, *class d* (delete) words not useful for the navigation window, and *class u* (uncertain) words whose usefulness in the navigation window was either neutral or difficult to judge. In the classification process, a judge used the *DualNAVI* system and examined the informativeness of each word as guidance. Classification into class *d* words was done conservatively because the consequences of removing informative words from the window are more serious than those of allowing useless words to appear.

Table 1 shows part of the classification of the 2,000 words. Words marked "*p*" are proper nouns. The difference between proper nouns in *class a* and proper nouns in other classes is that the former are wellknown. Most words classified as "*d*" are very common verbs (such as する(*do*) and 持つ(*have*)), adverbs, demonstrative pronouns, conjunctions, and numbers. It is therefore impossible to define a stop-word list by only using parts-of-speech because almost all parts-of-speech appear in *class d* words.

4.1.2 Measures used in the experiments

To evaluate the effectiveness of several measures, we compared the ability of each measure to gather (avoid) representative (non-representative) terms. We randomly sorted the 20,000 words and then compared the results with the results of sorting by other criteria: *Rep*(•, *LLR*), *Rep*(•, *DIFFNUM*), *tf* (term frequency), and *tf-idf*. The comparison was done by using the accumulated number of words marked by a specified class that appeared in the first *N* (1 ≤ *N* ≤ 2,000) words. The definition we used for *tf-idf* was

$$tf - idf = \sqrt{TF(T)} \times \log \frac{N_{total}}{N(T)},$$

where *T* is a term, *TF*(*T*) is the term frequency of *T*, N_{total} is the number of total documents, and *N*(*T*) is the number of documents that contain *T*.

4.1.3 Results

Figure 4 compares, for all the sorting criteria, the

accumulated number of words marked "*a*". The total number of *class a* words was 911. *Rep*(•, *LLR*) clearly outperformed the other measures. Although *Rep*(•, *DIFFNUM*) outperformed *tf* and *tf-idf* up to about the first 9,000 monograms, it otherwise under-performed them. If we use the threshold value of *Rep*(•, *LLR*), from the first word to the 1,511th word is considered representative. In this case, the recall and precision of the 1,511 words against all *class a* words were 85% and 50%, respectively. When using *tf-idf*, the recall and precision of the first 1,511 words against all *class a* words were 79% and 47%, respectively (note that *tf-idf* does not have a clear threshold value, though).

Although the degree of out-performance by *Rep*(•, *LLR*) is not seemingly large, this is a promising result because it has been pointed out that, in the related domains of term extraction, existing measures hardly outperform even the use of frequency (for example, Daille et al. 1994, Caraballo et al. 1999) when we use this type of comparison based on the accumulated numbers.

Figure 5 compares, for all the sorting criteria, the accumulated number of words marked by *d* (454 in total). In this case, fewer the number of words is better. The difference is far clearer in this case: *Rep*(•, *LLR*) obviously outperformed the other measures. In contrast, *tf-idf* and frequency barely outperformed random sorting. *Rep*(•, *DIFFNUM*) outperformed *tf* and *tf-idf* until about the first 3,000 monograms, but under-performed otherwise.

Figure 6 compares, for all the sorting criteria, the accumulated number of words marked *ap* (acceptable proper nouns, 216 in total). Comparing this figure with Fig. 4, we see that the out-performance of *Rep*(•, *LLR*) is more pronounced. Also, *Rep*(•, *DIFFNUM*) globally outperformed *tf* and *tf-idf,* while the performance of *tf* and *tf-idf* were nearly the same or even worse than with random sorting.

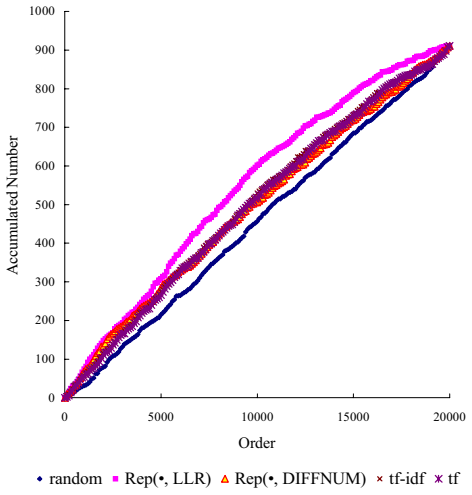

Figure 4
Sorting results on *class a* words

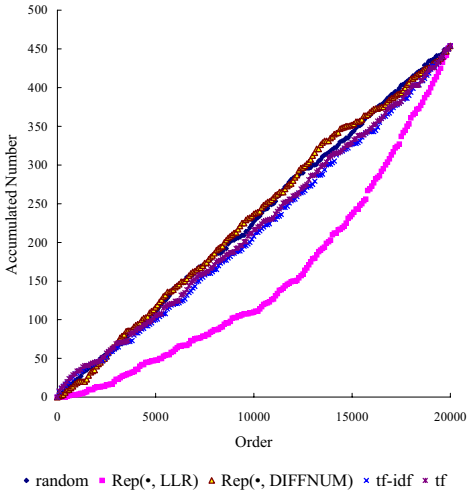

Figure 5
Sorting results on *class d* words

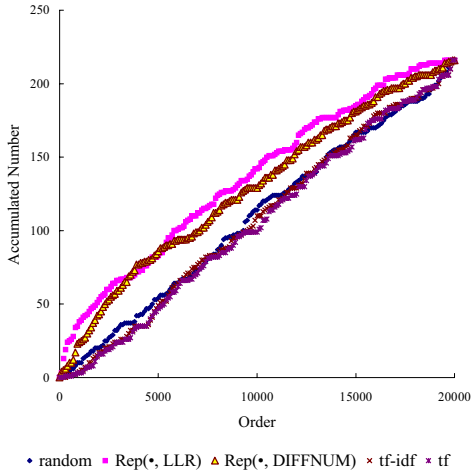

Figure 6
Sorting results on *class ap* words

Table 1
Examples of the classified words

class a	*class u*	*class d*
アミューズメントパーク (amusement park)	ひんやり (chilly)	八千三百万 (83,000,000)
脅迫状 (thretening letter)	消沈 (depressed)	多大な (greatly)
ファイアウオール (firewall)	石神 (Ishigami) p	千百四十六 (1, 146)
骨董品 (antique)	繁幸 (Shigeyuki) p	すべて (all)
アトランタ (Atlanta) p	筋違いだ(misdirected)	少しも (not…in the least)
	敏捷 (agility)	

In the experiments, proper nouns generally have a high *Rep*-value, and some have particularly high scores. Proper nouns having particularly high scores are, for instance, the names of *sumo* wrestlers or horses. This is because they appear in articles with special formats such as sports reports.

We attribute the difference of the performance between *Rep*(•, *LLR*) and *Rep*(•, *DIFFNUM*) to the quantity of information used. Obviously information on the distribution of words in a document is more comprehensive than that on the number of different words. This encourages us to try other measures of document properties that incorporate even more precise information.

4.2 Picking out frequent non-representative monograms

When we concentrate on the most frequent terms, *Rep*(•, *DIFFNUM*) outperformed *Rep*(•, *LLR*) in the following sense. We marked "clearly non-representative terms" in the 2,000 most frequent monograms, then counted the number of marked terms that were assigned *Rep*-values smaller than the threshold value of a specified representativeness measure.

The total number of checked terms was 563, and 409 of them are identified as non-representative by *Rep*(•, *LLR*). On the other hand, *Rep*(•, *DIFFNUM*) identified 453 terms as non-representative.

4.3 Rank correlation between measures

We investigated the rank-correlation of the sorting results for the 20,000 terms used in the experiments described in subsection 4.1. Rank correlation was measured by Spearman's method and Kendall's method (see Appendix) using 2,000 terms randomly selected from the 20,000 terms. Table 2 shows the correlation between *Rep*(•, *LLR*) and other measures. It is interesting that the ranking by *Rep*(•, *LLR*) and that by *Rep*(•, *DIFFNUM*) had a very low correlation, even lower than with *tf* or *tf-idf*. This indicates that a combination of *Rep*(•, *LLR*) and *Rep*(•, *DIFFNUM*) should provide a strong discriminative ability in term classification; this possibility deserves further investigation.

Table 2
Two types of Rank correlation between term-rankings by *Rep*(•, *LLR*) and other measures.

	Rep(•, *DIFFNUM*)	*tf-idf*	*tf*
Spearman	-0.00792	0.202	0.198
Kendall	-0.0646	0.161	0.153

4.4 Portability of baseline functions

We examined the robustness of the baseline functions; that is, whether a baseline function defined from a corpus can be used for normalization in a different corpus. This was investigated by using *Rep*(•, *LLR*) with seven different corpora. Seven baseline functions were defined from seven corpora, then were used for normalization for defining *Rep*(•, *LLR*) in the corpus used in the experiments described in subesction 4.1. The performance of the *Rep*(•, *LLR*)s defined using the different baseline functions was compared in the same way as in the subsection 4.1. The seven corpora used to construct baseline functions were as follows:

NK96-ORG: 15,8000 articles used in the experiments in 4.1
NK96-50000: 50,000 randomly selected articles from the whole corpus NK96 (206,803 articles of *Nikkei-shinbun* 1996)
NK96-100000: 100,000 randomly selected articles from NK96
NK96-200000: 200,000 randomly selected articles from NK96
NK98-158000: 158,000 randomly selected articles from articles in *Nikkei-shinbun* 1998
NC-158000: 158,000 randomly selected abstracts of academic papers from NACSIS corpus (Kando et al. 1999)
NC-ALL: all abstracts (333,003 abstracts) in the NACSIS corpus.

Statistics on their content words are shown in Table 3.

Table 3
Corpora and statistics on their content words

	NK96-ORG	NK96-50000	NK96-100000	NK96-200000
# of total words	42,555,095	13,498,244	26,934,068	53,816,407
# of different words	210,572	127,852	172,914	233,668

	NK98-158000	NC-158000	NC-ALL
# of total words	39,762,127	30,770,682	64,806,627
# of different words	196,261	231,769	350,991

Figure 7 compares, for all the baseline functions, the accumulated number of words marked "*a*" (see subsection 4.1). The performance decreased only slightly when the baseline defined from NC-ALL was used. In other cases, the differences was so small that they were almost invisible in Fig. 7. The same results were obtained when using *class d* words and *class ap* words.

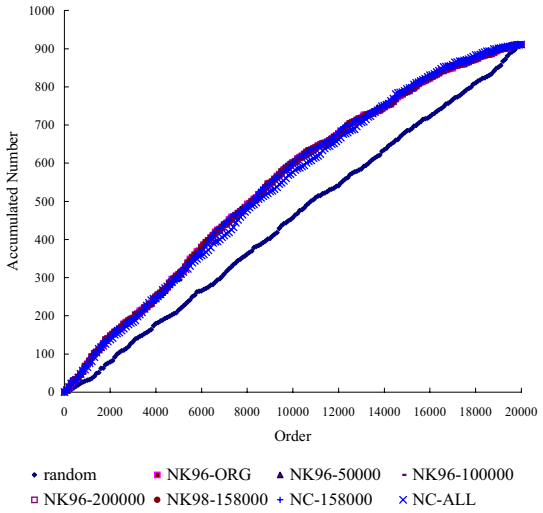

Figure 7
Sorting results on *class a* words

We also examined the rank correlations between the ranking that resulted from each representativeness measure in the same way as described in subsection 4.2 (see Table 4). They were close to 100% except when combining the Kendall's method and NACSIS corpus baselines.

Table 4
Rank correlation between the measure defined by an NK96-ORG baseline and ones defined by other baselines (%)

	NK96-50000	NK96-100000	NK96-200000	NK98-158000	NC-158000	NC-ALL
Spearmann	0.997	0.997	0.996	0.999	0.912	0.900
Kendall	0.970	0.956	0.951	0.979	0.789	0.780

These results suggest that a baseline function constructed from a corpus can be used to rank terms in considerably different corpora. This is particularly useful when we are dealing with a corpus similar to a known corpus but do not know the precise word distributions in the corpus. The same kind of robustness was observed when we used *Rep*(•,

DIFFNUM). This baseline function robustness is an important feature of measures defined using the baseline based.

5. Conclusion and future works

We have developed a better method -- the baseline method -- for defining the representativeness of a term. A characteristic value of all documents containing a term *T, D*(*T*), is normalized by using a baseline function that estimates the characteristic value of a randomly chosen document set of the same size as *D*(*T*). The normalized value is used to measure the representativeness of the term *T,* and a measure defined by the baseline method offers several advantages compared to classical measures: (1) its definition is mathematically simple and clear, (2) it can compare high-frequency terms with low-frequency terms, (3) the threshold value for being representative can be defined systematically, and (4) it can be applied to *n*-gram terms for any *n*.

We developed two measures: one based on the normalized distance between two word distributions (*Rep*(•, *LLR*)) and another based on the number of different words in a document set (*Rep*(•, *DIFFNUM*)). We compared these measures with two classical measures from various viewpoints, and confirmed that *Rep*(•, *LLR*) was superior. Experiments showed that the newly developed measures were particularly effective for discarding frequent but uninformative terms. We can expect that these measures can be used for automated construction of a stop-word list and improvement of similarity calculation of documents.

An important finding was that the baseline function is portable; that is, one defined on a corpus can be used for normalization in a different corpus even if the two corpora have considerably different sizes or are in different domains. We can therefore apply the measures in a practical application when dealing with multiple similar corpora whose word distribution information is not fully known but we have the information on one particular corpus.

We plan to apply *Rep*(•, *LLR*) and *Rep*(•, *DIFFNUM*) to several tasks in IR domain, such as the construction of a stop-word list for indexing and term weighting in document-similarity calculation.

It will also be interesting to theoretically estimate the baseline functions by using fundamental parameters such as the total number of words in a corpus or the total different number in the corpus. The natures of the baseline functions deserve further study.

Acknowledgements

This project is supported in part by the Advanced Software Technology Project under the auspices of Information-technology Promotion Agency, Japan (IPA).

References

Caraballo, S. A. and Charniak, E. (1999). Determining the specificity of nouns from text. *Proc. of EMNLP'99*, pp. 63-70.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography, *Computational Linguistics* 6(1), pp.22-29.

Daille, B. and Gaussier, E., and Lange, J. (1994). Towards automatic extraction of monolingual and bilingual terminology. *Proc. of COLING'94*, pp.515-521.

Dunning, T. (1993). Accurate Method for the Statistics of Surprise and Coincidence, *Computational Linguistics* 19(1), pp.61-74.

Firth, J. A synopsis of linguistic theory 1930-1955. (1957). *Studies in Linguistic Analysis*, Philological Society, Oxford.

Frantzi, K. T., Ananiadou, S., and Tsujii, J. (1996). Extracting Terminological Expressions, *IPSJ Technical Report of SIGNL*, NL112-12, pp.83-88.

Hisamitsu, T., Niwa, Y., and Tsujii, J. (1999). Measuring Representativeness of Terms, *Proc. of IRAL'99,* pp.83-90.

Kageura, K. and Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology* 3(2), pp.259-289.

Kando, N., Kuriyama, K., and Nozue, T. (1999). NACSIS test collection workshop (NTCIR-1), *Proc. of the 22nd Annual International ACM SIGIR Conf. on Research and Development in IR*, pp.299-300.

Kita, Y., Kato, Y., Otomo, T., and Yano, Y. (1994). Comparative Study of Automatic Extraction of Collocations from Corpora: Mutual Information vs. Cost Criteria, *Journal of Natural Language Processing,* 1(1), 21-33.

Nagao, M., Mizutani, M., and Ikeda, H. (1976). An Automated Method of the Extraction of Important Words from Japanese Scientific Documents, *Trans. of IPSJ*, 17(2), pp.110-117.

Nakagawa, H. and Mori, T. (1998). Nested Collocation and Compound Noun For Term Extraction, *Proc. of Computerm'98*, pp.64-70

Nishioka, S., Niwa, Y., Iwayama, M., and Takano, A. (1997). *DualNAVI*: An information retrieval interface. *Proc. of WISS'97*, pp.43-48. (in Japanese)

Niwa, Y., Nishioka, S., Iwayama, M., and Takano, A. (1997). Topic graph generation for query navigation: Use of frequency classes for topic extraction. *Proc. of NLPRS'97*, pp.95-100.

Noreault, T., McGill, M., and Koll, M. B. (1977). A Performance Evaluation of Similarity Measure, Document Term Weighting Schemes and Representation in a Boolean Environment. In Oddey, R. N. (ed.), *Information Retrieval Research*. London: Butterworths, pp.57-76.

Salton, G. and Yang, C. S. (1973). On the Specification of Term Values in Automatic Indexing. *Journal of Documentation* 29(4), pp.351-372.

Sparck-Jones, K. (1973). Index Term Weighting. *Information Storage and Retrieval* 9(11), pp.616-633.

Teramoto, Y., Miyahara, Y., and Matsumoto, S. (1999). Word weight calculation for document retrieval by analyzing the distribution of co-occurrence words, *Proc. of the 59th Annual Meeting of IPSJ*, IP-06. (in Japanese)

Appendix

Asusume that items $I_1,\ldots,I_N$ are ranked by measures A and B, and that the rank of item I_j assigned by A (B) is $R_A(j)$ ($R_B(j)$), where $R_A(i) \neq R_A(j)$ ($R_B(i) \neq R_B(j)$) if $i \neq j$. Then, Spearman's rank correlation between the two rankings is given as

$$1 - \frac{6 \times \sum_j (R_A(j) - R_B(j))^2}{N(N^2-1)},$$

and Kendall's rank correlation between the two rankings is given as

$$\frac{1}{{}_N C_2} \times \left(\{ \#\{(i,j) \mid \sigma(R_A(i) - R_A(j)) = \sigma(R_B(i) - R_B(j))\} - \#\{(i,j) \mid \sigma(R_A(i) - R_A(j)) = -\sigma(R_B(i) - R_B(j))\} \right),$$

where $\sigma(x)=1$ if $x > 0$, else if $x < 0$, $\sigma(x) = -1$.