HKIB-20000 & HKIB-40075: Hangul Benchmark Collections for Text Categorization Research

Jinsuk Kim, Ho-Seop Choe and Beom-Jong You
Department of Information Technology Research, KISTI, Daejeon, Korea
{jinsuk, hschoe, ybj}@kisti.re.kr

Jeong-Hyun Seo
Department of Cyber Environment Development, KISTI, Daejeon, Korea
jerry@kisti.re.kr

Suk-Hoon Lee
Department of Information & Statistics, Chungnam National University, Daejeon, Korea
shlee@stat.cnu.ac.kr

Dong-Yul Ra
Computer & Telecommunication Engineering Division,
Yonsei University, Wonju, Gangwon-do, Korea
dyra@yonsei.ac.kr

Received 16 June 2009; Revised 26 August 2009; Accepted 8 September 2009

The HKIB, or Hankookilbo, test collections are two archives of Korean newswire stories manually categorized with semi-hierarchical or hierarchical category taxonomies. The base newswire stories were made available by the Hankook Ilbo (The Korea Daily) for research purposes. At first, Chungnam National University and KISTI collaborated to manually tag 40,075 news stories with categories by semi-hierarchical and balanced three-level classification scheme, where each news story has only one level-3 category (*single-labeling*). We refer to this original data set as HKIB-40075 test collection. And then Yonsei University and KISTI collaborated to select 20,000 newswire stories from the HKIB-40075 test collection, to rearrange the classification scheme to be fully hierarchical but unbalanced, and to assign one or more categories to each news story (*multi-labeling*). We refer to this modified data set as HKIB-20000 test collection. We benchmark a k-NN categorization algorithm both on HKIB-20000 and on HKIB-40075, illustrating properties of the collections, providing baseline results for future studies, and suggesting new directions for further research on Korean text categorization

problem.

Categories and Subject Descriptors: Information Retrieval and Visualization [**Text Categorization, Test Collection**]

General Terms: Text Categorization, Korean Test Collection, Document Classification

Additional Key Words and Phrases: HKIB-20000, HKIB-40075, Hankook Ilbo, Hangul test collections, Machine learning, Newswire story, Evaluation, KRISTAL-IRMS, *k*-NN classifier, Multilabel, Hierarchical classification scheme

1. INTRODUCTION

Text categorization is an automatic task of assigning one or more predefined categories to a test document based on its content. While it is an active and mature research field in information retrieval and machine learning, text categorization research has been usually based on test collections of English texts such as Reuters-21578 [Lewis 1992; Lewis 2004], OHSUMED [Hersh et al. 1994], and Reuters Corpus Volume 1 (RCV1) [Lewis et al. 2004]. There also exists a test collection of multilingual news stories from Reuters, Ltd., released in the name of Reuters Corpus Volume 2 (RCV2) which contains over 487,000 Reuters News stories in thirteen languages[1].

These corpora are collections of documents to which human indexers have assigned categories from a predefined set. Test collections enable researchers to test ideas without hiring human indexers, and ideally to objectively compare results with published studies [Lewis et al. 2004]. However, people who are concerned with Korean text categorization might notice that, whereas extensive studies have been done to English document categorization, relatively few results have been reported on text categorization for Korean texts. This is mainly due to lack of proper Korean text collection designed particularly for text categorization research.

In this paper we introduce the Hankookilbo collection[2] which is a set of Korean text categorization test collections built on Korean newswire stories, of which stories are manually assigned with categories according to three-level hierarchical classification scheme. This collection consists of two versions of test collections, HKIB-40075 data set and HKIB-20000 data set. The HKIB-40075 test collection is single-labeled with leaf node categories (i.e., level-3 categories only). The HKIB-20000 is multi-labeled with non-leaf and/or leaf node categories (i.e., level 1, 2, and/or 3 categories).

One of the major differences between the two test collections is the category taxonomy. The HKIB-40075 is built on semi-hierarchical[3] category taxonomy while the HKIB-20000 was built on fully hierarchical category taxonomy (see Table I). Each news story in the HKIB-40075 test collection has only one level-3 category. This

[1]Dutch, French, German, Chinese, Japanese, Russian, Portuguese, Spanish, Latin American Spanish, Italian, Danish, Norwegian, and Swedish

[2]The Hankookilbo test collection is avaliable from the URL at http://www.kristalinfo.com/TestCollections/#hkib in a tarred and gzipped file. In addition, more detail characteristics of the two test collections are described in two on-line supplimentary materials, one [Kim 2009a] in Korean and the other [Kim 2009b] in English.

[3]Though it is described as semi-hierarchical, the category taxnomy used in the HKIB-40075 test collection is balanced at level-3 categories, and thus it can be regarded as *plain* category taxnomy.

Table I. Brief Description of the HKIB-40075 and HKIB-20000 Test Collections

Data Set	HKIB-40075	HKIB-20000
No. of documents	40,075	20,000
Category taxonomy	semi-hierarchical	hierarchical
Labeling policy	Single-label	multi-labels
Category levels assigned	level-3 (balanced)	level-1, 2, and/or 3 (unbalanced)
Total no. of categories assigned	40,075	23,434

characterizes the HKIB-40075 data set as a balanced and plain category taxonomy-based test collection, although the category set itself is hierarchical. However, as shown in many real field directory services, practical classification problems contain multi-label categorization and fully hierarchical category taxonomy. Since the HKIB-40075 data set is not suitable for researches that solve those practical problems, we changed the coding policies from single-label policy to multi-labels policy and from leaf-category policy to any-category policy while building the HKIB-20000 test collection.

Regarding big difference between English and Korean, it is interesting to see whether methods such as the SVM and k-NN that are effective for English text categorization problem are able to achieve similar performance for Korean by applying to the Hangul corpora presented in this paper.

We begin in Sections 2 and 3 by introducing the HKIB-40075 data set and the HKIB-20000 data set, respectively. Section 4 gives the design of evaluation of our variation of k-NN classifier on both test collections. Section 5 presents the benchmark results and observations. We end in Section 6 with some thoughts on further quality control and research directions that these two Hangul collections may support.

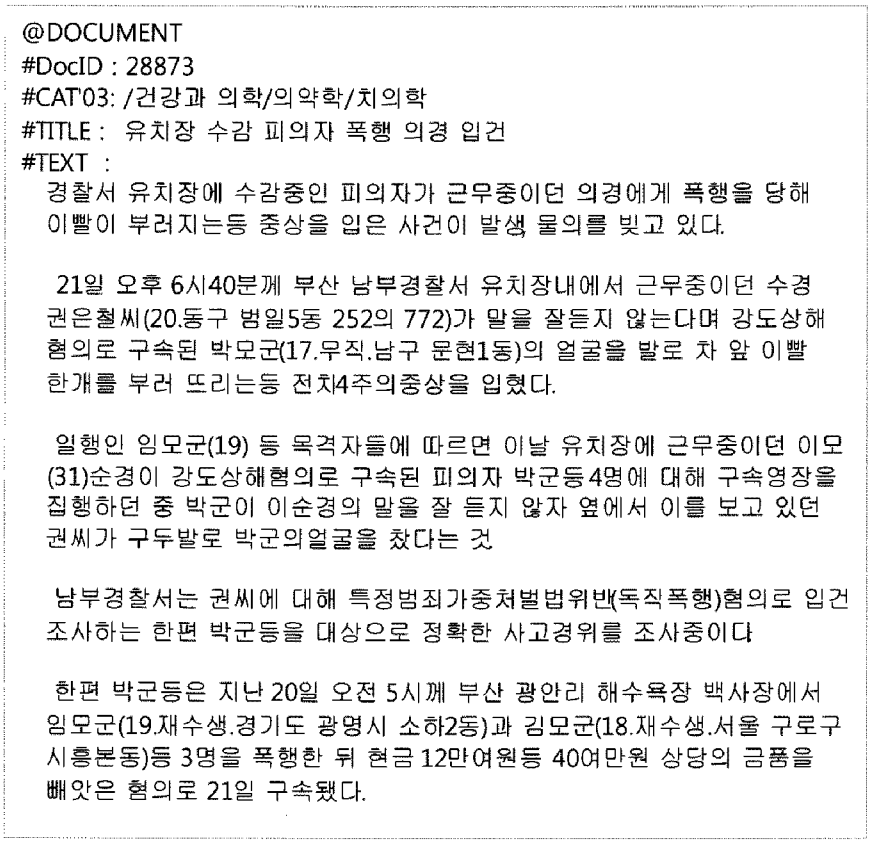
2. HKIB-40075 TEST COLLECTION

The HKIB-40075 test collection consists of 40,075 documents which is all Hankook-Ilbo news stories published from January 1, 1998 to December 31, 1999. Each document in this collection was manually assigned with only one leaf category. This collection is characterized to be a set of documents single-labeled with leaf node categories (level-3 categories).

2.1 Documents of HKIB-40075

For research purposes, the Hankook Ilbo (The Korea Daily), one of major newspaper companies in Korea, provided 40,075 Korean language news stories produced by its journalists between January 1, 1998 and December 31, 1999. The stories cover the range of content typical of a Korean language newswire of which length varies from a few dozens to several thousands of words. Figure 1 shows an example document from the HKIB-40075 data set.

The documents of HKIB-40075 test collection are distributed as five UTF-8 encoded text files each of which contains about one fifths of 40,075 news stories. Each story consists of a document delimiter(@DOCUMENT), a document identifier(#DocID :), a



@DOCUMENT
#DocID : 28873
#CAT'03: /건강과 의학/의약학/치의학
#TITLE : 유치장 수감 피의자 폭행 의경 입건
#TEXT :
경찰서 유치장에 수감중인 피의자가 근무중이던 의경에게 폭행을 당해 이빨이 부러지는등 중상을 입은 사건이 발생, 물의를 빚고 있다.

21일 오후 6시40분께 부산 남부경찰서 유치장내에서 근무중이던 수경 권은철씨(20.동구 범일5동 252의 772)가 말을 잘듣지 않는다며 강도상해 혐의로 구속된 박모군(17.무직.남구 문현1동)의 얼굴을 발로 차 앞 이빨 한개를 부러 뜨리는등 전치4주의중상을 입혔다.

일행인 임모군(19) 등 목격자들에 따르면 이날 유치장에 근무중이던 이모(31)순경이 강도상해혐의로 구속된 피의자 박군등4명에 대해 구속영장을 집행하던 중 박군이 이순경의 말을 잘 듣지 않자 옆에서 이를 보고 있던 권씨가 구두발로 박군의얼굴을 찼다는 것

남부경찰서는 권씨에 대해 특정범죄가중처벌법위반(독직폭행)혐의로 입건 조사하는 한편 박군등을 대상으로 정확한 사고경위를 조사중이다

한편 박군등은 지난 20일 오전 5시께 부산 광안리 해수욕장 백사장에서 임모군(19.재수생.경기도 광명시 소하2동)과 김모군(18.재수생.서울 구로구 시흥본동)등 3명을 폭행한 뒤 현금 12만여원등 40여만원 상당의 금품을 빼앗은 혐의로 21일 구속됐다.

Figure 1. An Example HKIB-40075 Document Written in Hangul (Korean alphabet).

manually assigned category(#CAT'03:)[4], title(#TITLE :) and body text(#TEXT :), which are formatted with nine-bytes long delimiters as shown below (See the example documents in Figures 1 and 2):

1. @DOCUMENT
   9 bytes long and single line. Starting point of a news story. This line should be discarded during parsing the text files.
2. #DocID :
   Single line. Document Identifier. This delimiter is followed by an identifier up to the end of the line. Each news story has a unique identifier. Documents with same DocID can be regarded as duplicated stories.
3. #CAT'03:
   Single line. This line contains only one category that tagged in the HKIB-40075 test collection. This delimiter is just followed by a three-level category up to the end of the line.
4. #CAT'07:
   Single line. This line contains one or more categories that tagged in the HKIB-20000 test collection. This delimiter is just followed by one or more categories up to the end of the line which is separated by semicolons(;). Refer Section 3 and

[4]In the case of HKIB-20000 test collection, additional category section is given to each news story at the line starting with #CAT'07: delimiter.

Figure 2. *This delimiter is not used in the HKIB-40075 test collection.*

5. #TITLE :
   Single line. Title of the news story. This delimiter is just followed by the title of the news story to the end of the line.
6. #TEXT :
   Multiple lines. Content of the news story. This delimiter is just followed by a new line character. The lines from just after this line and up to just before the next @DOCUMENT line constitute content of the news story.

2.2 Categories for HKIB-40075

To each document of 40,075 news stories provided by the Korea Daily, Korea Institute of Science & Technology Information (KISTI) and Chungnam National University collaborated to manually assign only one category from predefined set. At this stage, three-level hierarchical classification scheme was applied to the manual tagging. We refer to this classification scheme as the 2003-categories or 2003-category set. It is kept balanced by requiring that all leaf node categories are at the same depth of 3.

The 2003-category set is designed to represent the topic or the major subjects of a news story. It is hierarchically organized in nine level-1 topic groups: *Health and Medicine*, *Economy*, *Science*, *Education*, *Culture and Religion*, *Social Issue*, *Industry*, *Leisure*, and *Politics*. Each leaf node category has level-1 category which is followed by level-2 which is again followed by level-3 category. These leaf node categories are used in manual tagging for each news story. In a document, a leaf node category is expressed as "/level-1/level-2/level-3." Check out the file *HKIB-40075/hkib40075-cat03-all.categories* for a full list of categories and number of news stories assigned in this collection. There are 9 level-1, 32 level-2, and 120 level-3 categories in the 2003-

Table II. Document Distribution of the HKIB-40075 and HKIB-20000 Test Collections According to Level-1 Categories2 Document Distribution of the HKIB-40075 and HKIB-20000 Test Collections According to Level-1 Categories

Level-1 category	HKIB-40075		HKIB-20000	
	No. of docs	percentage	No. of docs	percentage
Health and Medicine	523	1.3	344	1.7
Economy	7300	18.2	6725	33.6
Science	794	2.0	554	2.8
Education	680	1.7	589	2.9
Culture and Religion	3457	8.6	3144	15.7
Social Issue	5273	13.2	4179	20.9
Industry	4890	12.2	3162	15.8
Leisure	614	1.5	271	1.4
Politics	16,544	41.3	872	4.4
/	-	-	160	0.8
Sum	40,075	100	20,000	100

category set. The document frequencies of the level-3 categories vary from 12 occurrences for /Health and Medicine/Medicine/Veterinary Science to 2704 for /Social Issue/Public Order/Police and Prosecutory.

2.3 HKIB-40075 Coding Policy

There were two coding policies for the HKIB-40075 test collection.

(1) *Single-labeling policy*: Only one category per news story should be assigned manually.

(2) *Leaf node category policy*: Only leaf node categories should be assigned to each story. There is no story which is assigned directly to non-leaf node categories.

These two policies were applied to manual coding of each news story in the HKIB-40075 test collection. As a result of these coding policies, each news story in the HKIB-40075 test collection has only one level-3 category. Therefore the HKIB-40075 test collection is said to be based on balanced and plain category taxonomy, though the 2003-category set itself is hierarchical.

@DOCUMENT
#DocID : 6298
#CAT'03: /건강과 의학/의약학/질병(암외의질병)
#CAT'07: /건강과 의학/질병/암외의질병;/정치/정부부처/보건복지부
#TITLE : 광견병 비상..광견병우려자 집단발생
#TEXT :
광견병주의보가 발령된 가운데 광견병 바이러스에 감염된 개에 물린 광견병우려자가 18일 4명이 새로 발생, 방역당국에 비상이 걸렸다.

보사부는 지난 14일 경기도 연천군에 거주하는 김모씨(48.농업)가 미친개(광견)에 물린지 4일만인 이날 같은 연천군내에서 강모 어린이(생후 9개월)와 강원도 화천군에서 3명이 광견에 물린 사실이 확인돼 방역당국의 격리 치료를 받고 있다고 밝 혔다.

보사부는 이들이 광견에 물렸으나 아직 광견병 증세를 나타내고 있지 않다고 설명하고 발병에 대비한 진료를 실시중이라고 덧붙였다.

보사부는 광견병의 대거발생이 우려됨에 따라 개에 대한 광견병 예방 접종을 무료로 실시하는 농수산부 및 국방부, 국립보건원 등 관련부처 관계자와 학계 전문가들을 이날 오후 긴급소집, 이동모 보건국장 주재로 회의를 열고 광견을 죽이는 등의 확산방지대책을 마련해서 즉각 실시키로 했다.

보사부는 이와 함께 개를 비롯 집에서 기르는 가축에 광견병 예방접종을 하고 가축이나 야생동물 등에 물렸을 때는 즉시 가까운 보건소에 신고해 주도록 국민들에게 당부했다.

광견병이란 원인바이러스인 레이비스 바이러스를 보유한 야생 또는 사육 동물에 물렸을 때 침에 들어 있는 바이러스의 전파로 발생하며 보통2-8주간의 잠복기를 거쳐 두통, 발열, 피로 등을 나타내며 심하면 신체마비 증세와 함께 물을 마시는 것을 두려워해 공수병이라고도 불린다.

광견병 환자는 지난 84년 서울에서 1명이 발생한 이래 지금까지 없었다.

Figure 2. An Example HKIB-20000 Document Written in Hangul.

3. HKIB-20000 TEST COLLECTION

The HKIB-20000 test collection is a modified and upgraded version of the HKIB-40075 data set where 20,000 documents from the HKIB-40075's 40,075 documents were carefully selected as base documents.

The classification scheme was modified to be a practical level as shown in real field directory services, and policy of category assignment was changed to multi-labeling with any level of three-level hierarchical classification system. This collection is characterized to be a set of documents multi-labeled with leaf and/or non-leaf node categories (level-1, level-2, and/or level-3 categories).

3.1 Documents of HKIB-20000

Among the original data of 40,075 news stories included in the HKIB-40075 data set, twenty thousands of stories were selected for this new version of test collection. The document distribution according to level-1 categories is shown in Table II for both HKIB-40075 and HKIB-20000 test collections. *Politics* is the biggest level-1 category in the HKIB-40075 test collection. However, it is one of the smallest category groups in the HKIB-20000. This feature will help future researches analyze over-fitting for large groups and under-fitting for small groups.

Figure 2 shows an example news story from the HKIB-20000 data set. Since the HKIB-20000 test collection was built based on the HKIB-40075, each of documents also retains a 2003-category (see Figure 2). The document format is as described in Section 2.1. Compared with the HKIB-40075 data set, each news story in this data set has an additional line which starts with "#CAT'07:" and is followed by one or more categories of the 2007-category set up to the end of the line. If more than one categories are assigned, they are separated by semicolon(;) as shown in Figure 2.

3.2 Categories for HKIB-20000

In order to build this data set, KISTI and Yonsei University collaborated to manually assign one or more categories to each of 20,000 documents. Carefully examining the three-level classification scheme of the HKIB-40075 test collection and finding that many level-2 categories that overlap level-3 categories, we decided to delete the level-2 categories from the 2003-category set and rearranged the classification scheme to have hierarchical three-level but unbalanced tree that does not require that all leaf node categories are at the same depth. We refer to this classification scheme as 2007-categories or 2007-category set.

Check out the file *HKIB-20000/hkib20000-cat07-all.categories* for full list of categories and number of news stories assigned. There are 9 level-1[5], 93 level-2, and 60 level-3 categories in the 2007-category set. In the contrary to the fact that each story of the HKIB-40075 test collection has been assigned with only one leaf node

[5]In fact, there is one more level-1 category designated as "/". Its frequency in the whole HKIB-2000 test collection is 160. This category means that the documents assigned to "/" do not belong to any of level-1 categories, i.e., these are unclassifiable documents. We included these documents in the training set but excluded from the evaluation process by removing them from each test set.

category of the 2003-category set, one or more of any category from leaf node categories as well as non-leaf node categories can be assigned to each story in the HKIB-20000 test collection. In other words, a news story of this collection has one or more categories which can be level-1, level-2 or level-3 categories. The document frequencies of the leaf node categories vary from one occurrence for /Economy/Bank to 1953 occurrences for /Economy/Business/Domestic. Total category assignments in this collection are 23,434 resulting in about 1.17 categories per news story. Among 20,000 news stories, 16.5% (3295 documents) are assigned with two or more categories.

3.3 HKIB-20000 Coding Policy

There were two coding policies for the HKIB-20000 test collection.

(1) *Multi-labeling policy*: One or more categories per news story can be assigned.

(2) *Any level category policy*: Level-1 and level-2 categories in addition to level-3 categories can be assigned to a news story.

These two policies were applied to manual coding of each news story in the HKIB-20000 test collection. During manual coding for this collection, we ignored 2003-categories previously defined in the HKIB-40075 test collection and each category assignment was started from scratch. As a result, 49% of the documents have different categories from their inherited 2003-categories.

Each news story in the HKIB-40075 test collection has only one level-3 category. This characterizes the HKIB-40075 data set as a balanced and plain category taxonomy-based test collection, although the 2003-category set itself is hierarchical. However, as shown in many real field directory services, practical classification problems contain multi-label categorization and fully hierarchical category taxonomy. Since the HKIB-40075 data set is not suitable for researches to solve those practical problems, we changed the coding policies from single-label to multi-labels policy and from leaf-category to any category policy while building the HKIB-20000 test collection.

4. BENCHMARKING METHODS

An important part of the value of a machine learning data set is the availability of published benchmark results. Good benchmark results serve to ensure that apparently superior new methods are not being compared to artificially low baselines. To provide such baselines, we ran a k-NN (k-nearest neighbor) classifier [Kim and Kim 2004] on the HKIB-40075 and HKIB-20000 data sets.

4.1 Test/Training Split

To support five-fold cross validation, we arbitrarily split the HKIB test collections to five subsets where each subset is contained in a text file. These files are included in the distribution file as HKIB-20000/HKIB-20000_00[1-5].txt and HKIB-40075/HKIB-40075_00[1-5].txt [Kim 2009a]. Using each file as a test set and the others as the training set, five splits are made and therefore five-fold cross-validation is possible for text categorization algorithms. Here, we refer to each split as Split 1, Split 2, Split 3, Split 4, and Split 5 for both data sets. For more information, refer to [Kim 2009a].

4.2 Effectiveness Measures

To measure the effectiveness of our text classifier, we use the F_1 measure [van Rijsbergen 1979] which corresponds to the harmonic mean of precision and recall:

$$P = \frac{categories\ relevant\ and\ retrieved}{categories\ retrieved}$$

$$R = \frac{categories\ relevant\ and\ retrieved}{categories\ relevant}$$

$$F_1 = \frac{2P_t}{2P_t+P_f+N_f} = \frac{2RP}{R+P}$$

where P_t is the number of documents a system correctly assigns to the category (true positives), P_f is the number of documents a system incorrectly assigns to the category (false positives), N_f is the number of documents that belong to the category but which the system does not assign to the category (false negatives), P is the precision (i.e., $P_t/(P_t + P_f)$, and R is the recall rate (i.e., $P_t/(P_t + N_f)$.

A special point of F_1 measure where precision is equal to recall is called *precision and recall break-even point*, or simply *break-even point* (BeP), i.e., *BeP=P=R*. Theoretically, BeP is always less than or equal to F_1 measure at any point. Therefore BeP is usually used to compare the effectiveness among different kinds of text categorization methods [Yang 1999; Sebastiani 2002]. We present BeP as the categorization effectiveness. On the other hand, if it is unable to get BeP, we present F_1 values close to BeP, where the difference between precision and recall rate is less than 0.001 or 0.1%.

To measure effectiveness across a set of categories we use both the *macroaverage* F_1 measure (unweighted mean of effectiveness across all categories) and the *microaverage* F_1 measure (effectiveness computed from the sum of per-category contingency tables).

4.3 Evaluation for Non-leaf Node Categories

As mentioned above, each HKIB document is assigned with categories of three-level hierarchical classification scheme. Therefore evaluation can be performed not only on the leaf node categories but also on the non-leaf node categories. During evaluation of categorization effectiveness for non-leaf node categories, each leaf node category is abbreviated to the corresponding non-leaf node category by removing subnode(s). For example, the leaf or level-3 category /Science/Social Science/Linguistic is abbreviated to /Science/Social Science for evaluation on level-2 categories and to /Science for level-1 category evaluation.

4.4 *k*-NN Classifier

The *k*-NN (*k-nearest neighbor*) classifiers have consistently been strong performers in text categorization evaluations [Yang 1999; Yang and Liu 1999]. As an example-based classifier, the *k*-NN classifier has many similarities with information retrieval systems. The variant *k*-NN [Kim and Kim 2004] we used here was built on an open-source information retrieval & management system, KRISTAL-IRMS [Kim et al. 2007], which was developed by KISTI to manage and retrieve XML documents and semi-structured texts such as bibliographies, theses, and journal articles. To retrieve

k top-ranked documents, our k-NN classifier uses a vector-space similarity measure ($Sim(q,d)$) between test document q and training document d, which is defined to be

$$Sim(q,d) = \frac{1}{W_d} \sum_{t \in q \wedge d} (w_{q,t} \cdot w_{d,t}) \tag{1}$$

with:

$$w_{d,t} = \log(f_{d,t}+1) \cdot \log\left(\frac{N}{f_t+1}+1\right)$$

$$w_{q,t} = \log(f_{q,t}+1) \cdot \log\left(\frac{N}{f_t+1}+1\right)$$

$$W_d = \log\left(\sum_{t \in d} f_{d,t}\right)$$

where $f_{x,t}$ is the frequency of term t in the document x; N is the total number of training documents; f_t is the number of documents where the term t occurs more than once; $w_{x,t}$ means the weight of term t in document x; and W_d represents the length of document d. Equation 1 is an empirically derived TF·IDF form of traditional vector-based information retrieval schemes [Witten et al. 1999] which have been commonly used due to its robustness and simplicity.

4.5 Category Relevance Measure

To determine whether a document d belongs to a category $c_j \in C=\{c_1, c_2, ..., c_{|C|}\}$ our text classifier retrieves k training documents most similar to d and computes c_j's weight by adding up similarities between d and documents that are retrieved and belong to c_j; if the weight is large enough, the decision is taken to be positive and negative otherwise. The weight of category c_j for document d is called *category relevance score, Relcjd*, and can be computed as follows: [Yang et al. 2002]

$$\mathrm{Re}l(c_j, d) = \frac{\sum_{d' \in R_k(d) \cap D_j} Sim(d', d)}{\sum_{d' \in R_k(d)} Si,(d', d)} \tag{2}$$

where $R_k(d)$ is the set of k-nearest neighbors (top-ranked training documents from the 1st to the kth) of document d; D_j is the set of training documents assigned category c_j; and $Sim(d', d)$ is the document-document similarity obtained by Equation 1 in Section 4.4. The denominator in Equation 2 is the sum of similarities between test document d and all of k top-ranked documents. It is used as a normalization factor that makes the category relevance score range from 0.0 to 1.0. For each test document, categories with relevance scores equal to or greater than a given threshold are assigned to the test document.

4.6 Experimental Setting

In our experiments, the content of an input document to be classified was in a concatenated form with title and body text of each news story. Terms separated by space characters were extracted as features from documents. And then, since the Hankookilbo news stories are written in Hangul (Korean text), we applied a Hangul

morpheme analyzer to Korean terms and included them into feature pool. Hangul morpheme analyzer used in this experiment is a module component of KRISTAL-IRMS which is used as the base of our k-NN classifier.

Feature selection was applied according to terms's document frequencies (DF) in the training set. In previous research it was shown that DF is a simple, effective, and reliable thresholding measure for selecting features in text categorization [Yang and Pedersen 1997]. In our experiments, by varying DF ranges during feature selection, we chose minimal DF(DF_{min}) as 2 and maximal DF(DF_{max}) as 5% of the training set. DF_{max} was set to 800 and 1600 for the HKIB-20000 and HKIB-40075, respectively.

k was selected to be 10 in our k-NN classifier. Several k values such as 3, 5, 7, 10, 20, 30, 40, and 50 were tested for five-fold cross validation and usually k=10 showed the best performance in our k-NN classifier (data not shown).

5. BENCHMARKING RESULTS

Tables III and IV give micro-averaged and macro-averaged values of F_1 measure for five test/train splits of the HKIB-40075 data set and HKIB-20000 data set, respectively. The results for categorization effectiveness are shown for non-leaf node categories (Level-1 and 2) in addition to leaf node categories (Level-3). As described in Section 4.5, we varied the threshold to get the break-even point when evaluating category relevance for each category level. The evaluation of our k-NN classifier for the HKIB-20000 data set on the 2003-category set was also conducted but the results are reported in README file for the Hankookilbo test collection [Kim 2009a] and also in a supplementary document [Kim 2009b].

Micro-averaged measures are dominated by high frequency categories [Lewis et al. 2004]. For level-3 categories of the 2003-category set in the HKIB-40075 test collection, 10% of top frequency categories (12 of 120) occupy 41% of all category assignments (16,293 of 40,075). For level-2 categories of the 2003-category set in the HKIB-40075 data set, 9% of top frequency categories (3 of 32) occupy 46% of all

Table III. Effectiveness of Our k-NN Classifier for 5 Test/training Splits of the HKIB-40075 Test Collection Evaluated with the *2003-category set*. miF_1 is Micro-averaged F_1 Measure and maF_1 is Macro-averaged F_1. As Explained in Section 4.2, These F_1 Values are Precision and Recall Break-even points. The Number Followed by *Level* at the First Column Means the Depth of Categories in the Three-level Classification Scheme

Test/Train Split	Split 1	Split 2	Split 3	Split 4	Split 5
Level-1 miF_1	0.794306	0.791581	0.795543	0.804657	0.793743
Level-1 maF_1	0.650182	0.671559	0.667802	0.678257	0.664961
Level-2 miF_1	0.732087	0.732446	0.730241	0.738674	0.730022
Level-2 maF_1	0.540164	0.550234	0.544932	0.556471	0.563859
Level-3 miF_1	0.619404	0.625212	0.614179	0.629913	0.618881
Level-3 maF_1	0.495485	0.493899	0.489127	0.515153	0.511786

[6]Note that the HKIB-20000 data set contains multi-labeled news stories. Due to multi-labeling, the total number of category assignments are 23,434 in 20,000 documents.

Table IV. Effectiveness of Our k-NN Classifier for Five Test/training Splits of the HKIB-20000 Test Collection Evaluated with the *2007-category set*. Other Details are as in Table III

Test/Train Split	Split 1	Split 2	Split 3	Split 4	Split 5
Level-1 miF_1	0.788566	0.797560	0.788019	0.795152	0.812399
Level-1 maF_1	0.717253	0.716409	0.718674	0.742369	0.745079
Level-2 miF_1	0.692622	0.701345	0.691800	0.707538	0.721281
Level-2 maF_1	0.474352	0.490875	0.475111	0.478200	0.544887
Level-3 miF_1	0.637102	0.652837	0.643187	0.651534	0.677645
Level-3 maF_1	0.469789	0.474778	0.470635	0.484926	0.561387

category assignments (18,248 of 40,075). For level-3 categories of the 2007-category set in the HKIB-20000 test collection, 10% of top frequency categories (16 of 161) occupy 44% of all category assignments (16,293 of 23,434)[6]. For level-2 categories of the 2007-category set in the HKIB-20000 data set, 10% of top frequency categories (10 of 103) occupy 53% of all category assignments (12,497 of 23,434). For level-1 categories for both HKIB data sets, the dominance of high frequency categories is greatly reduced, probably due to the fact that even the most infrequent categories are assigned to hundreds of news stories. For HKIB-40075 data set, level-1 category frequency ranges from 523 to 16,544. For HKIB-20000 data set, it ranges from 294 to 8,164.

On the other hand, macro-averaging gives equal weight to each category and thus is dominated by low frequency categories [Lewis et al. 2004]. For both HKIB data sets, this tendency can be seen in Figures 3 and 4, which is stronger in level-2 and level-3 categories than in level-1. Regarding level-1 categories, the dominance of low frequency categories are greatly decreased in the HKIB-20000 data set, compared with HKIB-40075 (Compare Figure 3(a) and 4(a)).

Figure 3 shows F_1 values of KRISTAL's k-NN classifier on the level-1, level-2, and level-3 categories for the first Test/Train split of the HKIB-40075 test collection. The F_1 values are sorted by training set frequency of the category (designated as Category Frequency in Figure 3). Plots for the other four splits showed very similar pattern

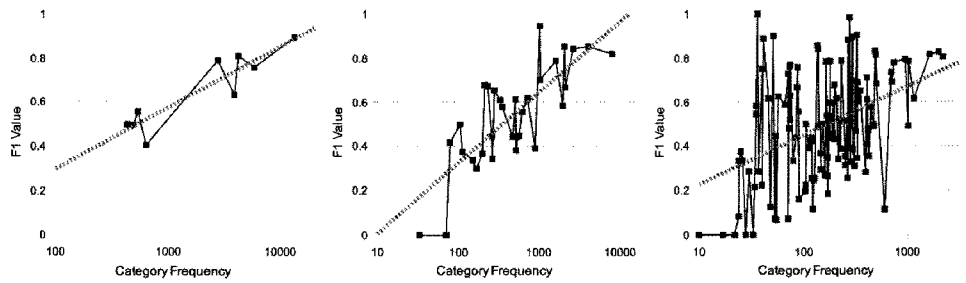


Figure 3. Test set F_1 of the HKIB-40075 Split 1 for our kNN classifier approach on level-1, 2, and 3 categories of which category counts in the test set are 9, 32, and 120, respectively. Categories are sorted by training set frequency, which is shown on the x-axis. A trend line is fitted for each plot as thick and dotted line. We set the threshold to get the micro-averaged break-even point for each category level.

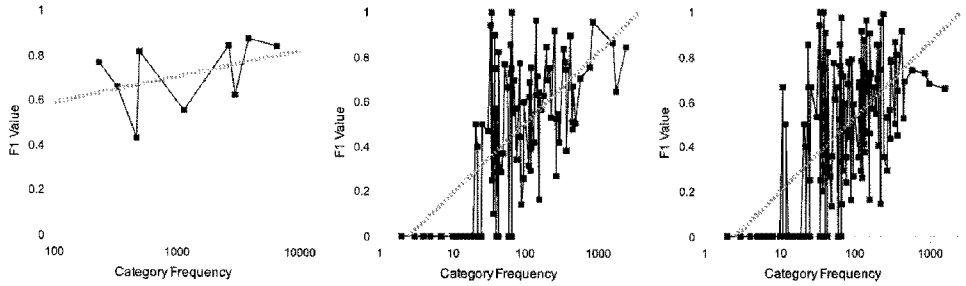


Figure 4. Test set F_1 of HKIB-20000 Split 1 for our kNN classifier approach on level-1, 2, and 3 categories of which category counts in the test set are 9, 93, and 147, respectively. Other details are as in Figure 3.

which are reported in a separate on-line document [Kim 2009b].

Figure 4 shows F_1 values for our k-NN classifier on the level-1, level-2, and level-3 categories sorted by category frequency for the first Test/Train split of the HKIB-20000 test collection. Plots for the other four splits showed very similar pattern which are reported in a supplementary on-line document [Kim 2009b].

As shown in Figures 3 and 4, effectiveness generally increases with increasing class frequency, but the category-to-category variation is very large, especially in the case of level-3 for HKIB-40075, and level-2 and level-3 for HKIB-20000. This variation seems to be correlated with the number of total categories as it can be seen that total category counts of 103 (Figure 4(b)), 120 (Figure 3(c)), and 161 (Figure 4(c)) show large variations, while 9 (Figure 3(a) and 4(a)) and 32 (Figure 3(b)) show relatively small variations.

More detail characteristics of the two test collections are described in two on-line materials, one [Kim 2009a] in Korean and the other [Kim 2009b] in English. We hope that these can help readers understand the quality and characteristics of the HKIB test collections.

6. DISCUSSION

Good benchmark results serve to ensure that apparently superior new methods are not being compared to artificially low baselines. We ran KRISTAL's k-NN classifier on the HKIB-40075 and HKIB-20000 data sets and provided such baselines. However, there are several other popular supervised learning approaches such as SVM (Support Vector Machine), Bayesian classifiers, decision tree, and Rocchio-style algorithms. Due to lack of appropriate Korean language processing tools for such algorithms we do not provide baselines for those algorithms. Future studies may include applying such algorithms to the Hankookilbo collection to provide baselines according to classifiers.

Korean researchers on text categorization have suffered from lack of test collections for text categorization built on Korean language corpus. While extensive studies have been done to English text categorization, few results have been reported on text categorization for Korean texts. In this paper we introduced the Hankookilbo collection which is a set of Korean text categorization test collections built on Korean

newswire stories. Using these corpora and taking into account the big difference between English and Korean, we hope that scientific and technological progress of categorization methods such as SVM and k-NN can be achieved for Korean language documents, as shown successfully for the English text categorization problem.

Applying techniques for Korean language processing such as Hangul morpheme analysis, complex noun extraction, and removal of common words in Korean (stopwords) to feature extraction and feature selection is still remained as future studies. We also believe that the Hankookilbo collection can support substantial research in hierarchical categorization, effectiveness of low frequency categories, sampling strategies, and other areas. Finally, we hope that our benchmark data will encourage replicability and transparency in Korean language text categorization research.

ACKNOWLEDGMENT

We would like to thank human indexers for their painful efforts to manually assign appropriate categories to the news stories of the HKIB-40075 and HKIB-20000 test collections. The first author would like to especially thank Changmin Kim and Jieun Chong for supporting this work.

REFERENCES

HERSH, W., C. BUCKLEY, T. J. LEONE, AND D. H. HICKMAN. 1994. OHSUMED: an Interactive Retrieval Evaluation and New Large Text Collection for Research. In *Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 94)*, 192–201.

KIM, JINSUK AND MYOUNG HO KIM. 2004. An Evaluation of Passage-based Text Categorization. *Journal of Intelligent Information Systems* 23(1):47–65.

KIM, JINSUK, DU-SEOK JIN, YUNSOO CHOI, CHANG-HOO JEONG, KWANGYOUNG KIM, SUNG-PIL CHOI, MINHO LEE, MIN-HEE CHO, HO-SEOP CHOE, HWA-MOOK YOON, AND JEONG-HYUN SEO. 2007. Toward DB-IR Integration: Per-Document Basis Transactional Index Maintenance. In *The 6th International Conference on Advanced Language Processing and Web Information Technology (ALPIT 2007)* 6:452–462, Luoyang, China.

KIM, JINSUK. 2009. HKIB-20000/HKIB-40075 Korean Text Categorization Test Collections. README file (version 1.0). Manuscript, May 31, 2009.

http://www.kristalinfo.com/TestCollections/readme_hkib.html

KIM, JINSUK. 2009. Experimental Results for KRISTAL's kNN Classifier on HKIB-20000 & HKIB-40075 Hangul Benchmark Collections for Korean Text Categorization Research. Manuscript, June 10, 2009.

http://www.kristalinfo.com/TestCollections/supp_hkib.pdf

LEWIS, DAVID D. 1992. An Evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 92)*, 37–50.

LEWIS, DAVID D. 2004. Reuters-21578 Text Categorization Test Collection. Distribution 1.0 README file (version 1.3). Manuscript, May 14, 2004.

http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt

LEWIS, DAVID D., YIMING YANG, TONY G. ROSE, AND FAN LI. 2004. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research* 5:361–397.

SEBASTIANI, F. 2002. Machine Learning in Automated Text Categorization. *ACM Computing Surveys* 34(1):1–47.

VAN RIJSBERGEN, C. J. 1979. *Information Retrieval.* Buttersworths, London, second edition.

WITTEN, I. H., A. MOFFAT, AND T. C. BELL. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images.* San Francisco: Morgan Kaufmann Publishing.

YANG, Y. AND J. O. PEDERSEN. 1997. A Comparative Study on Feature Selection in Text Categorization. In *The Fourteenth International Conference on Machine Learning (ICML 97)*, 412–420.

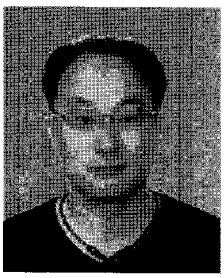
YANG, Y. 1999. An Evaluation of Statistical Approaches to Text Categorization. *Journal of Information Retrieval* 1(1):67–88.

YANG, Y. AND X. LIU. 1999. A Re-examination of Text Categorization Methods. In *Proceedings of the Twenty-Second Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 99)*, 42–49.

YANG, Y., S. SLATTERY, AND R. GHANI. 1999. A Study of Approaches to Hypertext Categorization. *Journal of Intelligent Information Systems* 17(2):219–241.

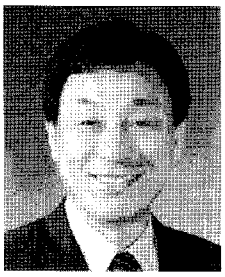
Jinsuk Kim is Senior Researcher of Department of Information Technology Research at Korea Institute of Science & Technology Information (KISTI), Republic Korea. He received M.Sci in Biology and M.Eng. in Computer Science from the Korea Advanced Institute of Science & Technology (KAIST), Republic of Korea. His research interests lie in information retrieval, automated text categorization, bioinformatics and DB-IR integration.

Ho-Seop Choe is Senior Researcher of Department of Information Technology Research at Korea Institute of Science & Technology Information (KISTI), Republic of Korea. He received Master degree in Korean language from the Kyeong-Nam University, Korea., and Ph.D. in Computer Science from Ulsan University, Korea, respectively. His research interests lie in the intelligent word network and knowledge base.

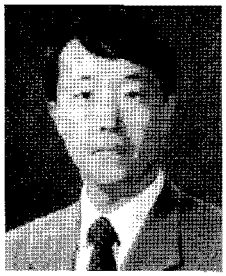


Beom-Jong You is Principal Researcher of Department of Information Technology Research at Korea Institute of Science & Technology Information (KISTI), Republic of Korea. He received Master and Ph.D. degrees in Library and Information Science from the Chungnam National University, Korea. His research interests lie in the information science, knowledge bases and semantic technologies.

Jeong-Hyun Seo is Senior Researcher of Department of Cyber Environment Development at Korea Institute of Science & Technology Information (KISTI), Republic of Korea. He received B.S. degree in mathematics from the Hanyang University, Korea. And he is currently a Master and Ph.D. candidate in the Computer Science at the Yonsei University, Korea. His research interests lie in the information retrieval and management and super-computing.



Suk-Hoon Lee is a professor of Department of Information & Statistics, Chungnam National University, Korea. He received the B.S. and Ph.D. degrees from Sogang University, Korea, and Ohio State University, USA, respectively. His research interests lie in statistics, statistical optimization, life testing data analysis, data mining and information retrieval.



Dong-Yul Ra is a professor of Computer & Telecommunication Engineering Division, Yonsei University, Korea. He received the B.S., M.S., and Ph.D. degrees from Seoul National University, KAIST, Korea, and Michigan State University, USA, respectively. His research interests lie in natural language processing, artificial intelligence and information retrieval.