Computational Linguistics and Chinese Language Processing
Vol. 5, No. 2, August 2000, pp. 1-42

Adaptive Word Sense Disambiguation Using Lexical Knowledge in a Machine-readable Dictionary

Jen Nan Chen*

Abstract

This paper describes a general framework for adaptive conceptual word sense disambiguation. The proposed system begins with knowledge acquisition from machine-readable dictionaries. Central to the approach is the adaptive step that enriches the initial knowledge base with knowledge gleaned from the partial disambiguated text. Once the knowledge base is adjusted to suit the text at hand, it is applied to the text again to finalize the disambiguation decision. Definitions and example sentences from the Longman Dictionary of Contemporary English are employed as training materials for word sense disambiguation, while passages from the Brown corpus and Wall Street Journal (WSJ) articles are used for testing. An experiment showed that adaptation did significantly improve the success rate. For thirteen highly ambiguous words, the proposed method disambiguated with an average precision rate of 70.5% for the Brown corpus and 77.3% for the WSJ articles.

Keywords: word sense disambiguation, machine-readable dictionary, semantics.

1. Introduction

Word sense disambiguation is a long-standing problem in natural language understanding. It seems to be very difficult to statistically acquire enough word-based knowledge about a language to build a robust system capable of automatically disambiguating senses in unrestricted text. For such a system to be effective, a large number of balanced materials must be assembled in order to cover many idiosyncratic aspects of the language. There exist three issues in a lexicalized statistical word sense disambiguation (WSD) model: data sparseness, the lack of abstraction, and static learning. First, a word-based model has a plethora of parameters that are difficult to estimate reliably even with a very large corpus. Under-trained models lead to low precision. Second, word-based models lack a degree of abstraction

* Department of Information Management, Ming Chuan University, Shih-lin, Taiwan, R.O.C

that is crucial for a broad coverage system. Third, a static WSD model is unlikely to be robust and portable, since it is very difficult to build a single model relevant to a wide variety of unrestricted texts. Several WSD systems have been developed that apply word-based models to a specific or genre domain to disambiguate senses appearing in generally easy context that has a large number of typically salient words. In the case of unrestricted text, however, the context tends to be very diverse and difficult to capture with a lexicalized model; therefore, a corpus-trained system is unlikely to transfer well to a new domain.

Generality and adaptability are, therefore, keys to a robust and portable WSD system. A concept-based model for WSD requires fewer parameters and has an element of generality built in. Conceptual classes make it possible to generalize from word-specific context in order to disambiguate word senses appearing in an unfamiliar context in terms of word recurrences. An adaptive system, armed with an initial lexical and conceptual knowledge base extracted from machine-readable dictionaries (MRD), has two strong advantages over static lexicalized models trained on a corpus. First, the initial knowledge is rich and unbiased enough for a substantial portion of text to be disambiguated correctly. Second, based on the result of initial disambiguation, an adaptation step can then be performed to make the knowledge base more relevant to the task at hand, thus resulting in broader and more precise WSD.

In this paper we explore in some depth the question of whether conceptual knowledge in the MRD is effective enough to provide a general solution for disambiguating contexts of unrestricted texts, such as the Brown and Wall Street Journal (WSJ) corpora. Major emphasis has previously been placed on self-adaptation [Chen and Chang 1998a]. This approach is based on the hypothesis that a substantial part of a given text is *easy* or prototypical and, therefore, susceptible to interpretation based on general knowledge derived from the MRD. By adapting the contextual representation of word senses to those in the easy context, we hope to be better equipped to interpret the other part, which is usually considered a *hard* context. Adaptation results in gaps in the general knowledge being filled in or domain specific information being added to the initial knowledge base. Either way, adaptation makes the knowledge base more relevant to the text and, therefore, more effective for WSD in a hard context. We will give experimental results showing the effectiveness of this adaptive WSD approach based on initial knowledge base acquired from the MRD. Although our adaptive approach requires virtually no domain-specific training, it nevertheless achieves high precision rates for WSD of unrestricted text rivaling those of static methods that demand very lengthy training using a very large corpus.

Figure 1 lays out the general framework for the adaptive conceptual WSD approach which this research employed. The learning process described here begins with a step involving knowledge acquisition from MRDs. With this acquired knowledge, the input text is read and a trial disambiguation step is carried out. An adaptation step follows which combines the initial knowledge base with knowledge gleaned from the partially disambiguated text. Once the knowledge base is adjusted to suit the text at hand, it is then applied to the text again to finalize the disambiguation result. For instance, the

initial contextual representation (CR) extracted from the Longman Dictionary of Contempory English [Proctor 1978, LDOCE] for the *bank*-GEO sense contains both lexical and conceptual information: {*land*, *river*, *lake*, ···} ∪ {GEO, MOTION, ···}. The initial CR is informative enough to disambiguate a passage containing "*a deer near the river bank*" in the input text. The trial disambiguation step produces sense tagging of *deer*/ANIMAL and *bank*/GEO, but certain instances of *bank* are left untagged due to the lack of WSD knowledge. We observe that the *bank*-GEO sense in the context of *vole* is unresolved since there is no link between ANIMAL and GEOGRAPHY. Subsequently, the adaptation step adds *deer* and ANIMAL to the contextual representation for *bank*-GEO. The adapted CR is now enriched with information capable of disambiguating the instance of *bank* in the context of *vole* to produce the final disambiguation result.

The rest of this paper is organized as follows. First of all, we will peresent how easy contexts are interpreted and ambiguous words are labeled in the initial disambiguation step using general knowledge derived from MRD. Next, we describe the adaptation step that uses the sense labels assigned to polysemous words. After that, we will describe the strategy of using the adapted knowledge base and defaults. Next, we will give a detailed account of experiments conducted to assess the effectiveness of the adaptive approach, including the experiment setup, results and evaluation. Following that, we will review the recent WSD literature from the perspective of various types of contextual knowledge and different representation schemes. Finally, we will draw conclusions.

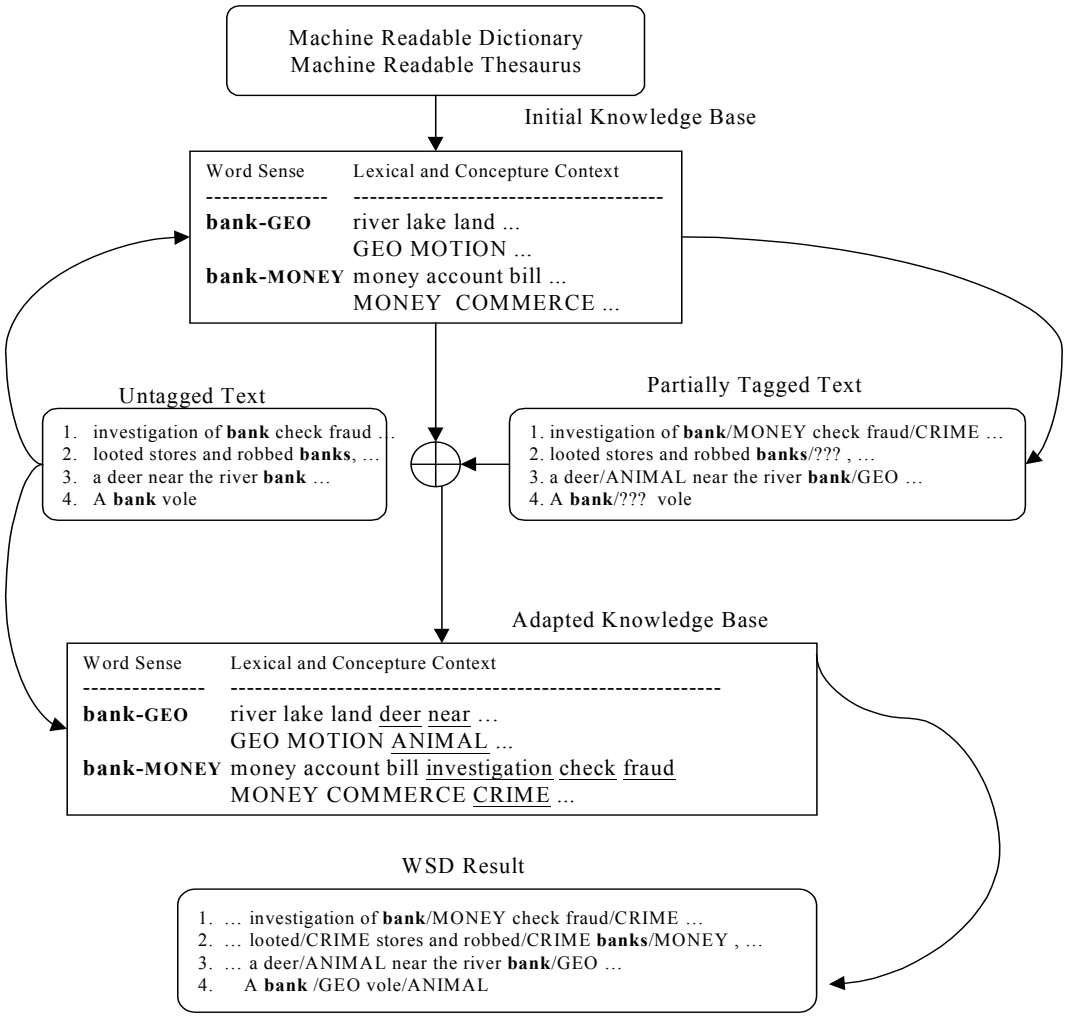


Figure 1 *General framework for adaptive WSD using MRD.*

2. Acquisition of Disambiguation Knowledge using MRD

In this section, we will describe how the conceptual characterization technique is applied to MRD definitions and give examples of acquiring WSD knowledge. First, we will show word level definitions based on a lexical CR and then a conceptual CR. Next, we will show the advantage of including information gained from an example sentence. Finally, we will combine these techniques to perform adaptative WSD computation.

2.1 Contextual Representation from Sense Definition

2.1.1 Lexicalized Contextual Representation

A word-level contextual representation from MRD definitions can be derived almost effortlessly. Let *CR*(*W*, *S*) denote the contextual representation of the sense *S* of headword *W*. Intuitively, it is composed of the content words in the definition with a specific sense. Thus, *CR*(*W*, *S*) can be represented symbolically as $\{ x \mid x \in DEF^{1}_{W,S}$ and x is not a function word $\}$. To illustrate, the CRs for the nine nominal senses of *bank* in LDOCE listed below are shown in Table 1:

bank.1.n.1	land along the side of a river, lake, etc.;
bank.1.n.2	earth which is heaped up in a field or garden, often making a border or division;
bank.1.n.3	a mass of snow, clouds, mud, etc.;
bank.1.n.4	a slope made at bends in a road or race-track, so that they are safer for cars to go round;
bank.1.n.5	a high underwater of bank in a river, harbour, etc.;
bank.3.n.1	a row, esp. of OAR in an ancient boat or KEY on a TYPEWRITER;
bank.4.n.1	a place in which money is kept and paid out on demand, and where related activities go on;
bank.4.n.2	a place where something is held ready for use, esp. ORGANIC products of human origin for medical use;
bank.4.n.3	a person who keeps a supply of money or pieces for payment or use in a game of chance.

Table 1. *Lexical contextual representations for* bank *senses.*

Sense ID	Sense Label *S*	Lexical Context Representation $LCR(D_{bank,s})$
bank.4.n.1	MONEY	{place, money, keep, pay, demand, activity}
bank.1.n.1	RIVER	{land, lake, river}
bank.1.n.5	SANDBANK	{underwater, sand, harbour}
bank.1.n.2	EARTH	{earth, heap, field, garden, boarder, division}
bank.1.n.3	PILE	{mass, snow, cloud, mud}
bank.1.n.4	ROAD	{car, aircraft, move, side, turn}
bank.3.n.1	ROW	{row, oar, boat, key, typewriter}
bank.4.n.2	MEDICINE	{place, hold, use, organic, product, human, origin, medical}
bank.4.n.3	GAMBLE	{person, keep, supply, money, payment, game, chance}

2.1.2 Conceptualized Contextual Representation

The word-based CR from MRD definitions is highly precise and effective but not broad enough to work alone effectively. Word-based sense representation is hampered by the difficulty of providing estimates for a very large parameter space leading to limited coverage in WSD. Certainly, there are many

[1] By $DEF_{S,W}$, we mean the definition of sense *S* of headword *W* in MRD.

situations that call for a conceptual generalization of a word-based representation of word sense from an example sentence. For instance, the RIVER sense of *bank* in Example (1c) can be correctly interpreted by an MRD-based CR, but only when the contextual word *river* in the CR is generalized to all words related to RIVER, including the word *stream*:

(1) a. a ribbon of mist along the river bank;
b. a small excavation in the river bank;
c. the left bank of the stream.

There are many possible approaches to making such a generalization and deriving a conceptualized CR (CCR) of word sense. Chen and Chang [1998b] described one such approach based on thesaurus topics. The CCR for each MRD sense can be viewed as relating to words listed under some Longman Lexicon of Contemporary English [McArthur 1992, LLOCE] topics. By linking MRD senses to thesaurus senses and by classifying senses according to linked senses, we can derive the CCR for a sense definition. Table 2 shows the topical CCR for the senses of *bank* in LDOCE. Each sense in MRD is given a list of weighted LLOCE topics. The weights in the CCR are normalized to a sum of unity for the obvious reason. Table 3 shows the lists of words listed in LLOCE under the topics relevant to *bank* senses.

Table 2. *Conceptual context representations for a definition* D *of* bank *senses.*

Sense Label *S*	Topics[2] (with weights) on $CCR(D_{bank, s})$
MONEY	Je(0.45), Jf(0.33), Jd(0.22)
RIVER	Ld(0.45), Mf(0.26), Me(0.14), Hc(0.07), Af(0.05), Ad(0.04)
EARTH	La(0.36), Ld(0.24), Eg(0.20), Me(0.12), Ie(0.08)
PILE	Lc(0.59), Db(0.13), Hc(0.09), La(0.09), Md(0.09)
ROAD	Md(0.45), Me(0.38), Ld(0.17)
ROW	Md(0.49), Gd(0.18), Mc(0.16), Kb(0.12), Me(0.06)
MEDICINE	Bd(0.70), Bj(0.30)
GAMBLE	Ke(0.35), Kh(0.28), Kf(0.23), Cn(0.14)

We sum up the above description and outline the procedure as Algorithm 1 for creating a CCR for a word *W* of sense *S* with definition *D*.

Algorithm 1: Creating a conceptualized contextual representation $CCR(D_{w,s})$ for a word *W* of sense *S* with definition *D*.

Step 1: Run the *TopSense* algorithm described by Chen and Chang [1998b] to map *D* to *SC(D)*, a set of semantic categories in a thesaurus.

[2] See Appendix A for more details.

Step 2: Create a conceptualized contextual representation $CCR(D_{w,s})$ for sense S with definition D:

$$CCR(D_{w,s}) = \sum_{T \in SC(D)} WORD(T)^{3},$$

where *WORD(T)* is a set of related words in semantic category *T*.

Table 3. *Definition-based conceptual context representation and related word lists for* bank *senses*

Sense Division *S*	Topics	Word List on $CCR(D_{bank,s})$
MONEY	Je(0.45) Jf(0.33) Jd(0.22)	{money, pay, cash, capital, account, charge, ... pay, bond, bill, charge, ... money, cash, fund, check, ... }
RIVER	Ld(0.45) Mf(0.26) Me(0.14) Hc(0.07) Af(0.05) Ad(0.04)	{lake, land, river, shore, stream, beach, ... boat, ship, craft, port, ... place, edge, road, border, ... rock, stone, clay, soil, ... fish, crab, coral, shell, fur, ... chicken, duck, goose, seabird, ...}
EARTH	La(0.36) Ld(0.24) Eg(0.20) Me(0.12) Ie(0.08)	{universe, space, planet, constellation, ... lake, land, river, shore, stream, beach, ... farming, field, crop, stock, productive, ... place, edge, road, border, ... playgroup, school, college, classroom, ...}
PILE	Lc(0.59) Db(0.13) Hc(0.09) La(0.09) Md(0.09)	{weather, climate, sky, cloud, fog, steam, ... roof, ceiling, wall, door, ground, ... rock, stone, clay, soil, ... universe, space, planet, constellation, ... transport, vehicle, car, motorcar, transit, ...}
ROAD	Md(0.45) Me(0.38) Ld(0.17)	{transport, vehicle, car, motorcar, transit, ... place, edge, road, border, ... lake, land, river, shore, stream, beach, ...}
ROW	Md(0.49) Gd(0.18) Mc(0.16) Kb(0.12) Me(0.06)	{transport, vehicle, car, motorcar, transit, ... printing, sign, letter, code, … sail, caravan, itinerary, … song, melody, dance, … road, street, …}
MEDICINE	Bd(0.70) Bj(0.30)	{blood, trunk, breast, back, buttock, waist, ... patient, examine, diagnose, soothe, ...}
GAMBLE	Ke(0.35) Kh(0.28)	{athletics, run, jump, ride, game, round, ... ball game, shoot, golf, pitch, football, ...

[3] *WORD*(*T*) is a bag of words rather than a set. By summation of bags, we mean collecting all the word instances in the bags and keeping track of counts.

Sense Division *S*	Topics	Word List on $CCR(D_{bank,s})$
MONEY	Je(0.45)	{money, pay, cash, capital, account, charge, ...
	Kf(0.23)	cards, pack, suit, heart, club, poker, dice, ...
	Cn(0.14)	war, warfare, conflict, fight, ...}

An Illustrative Example

In the following, we demonstrate how Algorithm 1 works. Given the sense definition of bank.4.n.1 shown in Section 2.1.1, the *CCR*(bank.4.n.1) can be acquired by means of Algorithm 1.

Step 1: After running the *TopSense* algorithm, we have *SC*(bank.4.n.1) = {**Je**, **Jf**, **Jd**}.

Step 2: Next, we expand each of three topics in *SC*(bank.4.n.1) to a cluster of words. Thus, we have

WORD (**Je**) ={money, pay, cash, capital, account, charge, ...},

WORD (**Jf**) ={pay, bond, bill, charge, ...} and

WORD (**Jd**) ={money, cash, fund, check, ...}.

Finally, the *CCR* (**bank.4.n.1**) =*WORD* (**Je**) + *WORD* (**Jf**) + *WORD* (**Jd**)
={money, pay, cash, capital, ...,
pay, bond, bill, charge, ...,
money, cash, fund, check, ... }.

2.2 Contextual Representation from an Example Sentence

Dictionary examples are intended to show typical use of words in context. Therefore, MRD examples provide rich information supplementary to definitions. In this section, we will describe a method for tagging bilingual sentences with sense labels based on dictionary definitions and translations in a bilingual MRD.

However, the sense for each word in an example is not explicitly marked except for the word being defined. That limits the potential for using dictionary examples as knowledge sources for WSD. Gale, Church and Yarowsky [1992b] first pointed out that the strong constraint of one-sense-per-translation can be exploited to tag a bilingual corpus for training a statistical WSD model. Building on their idea, we describe a new method for tagging bilingual sentences, in the MRD or elsewhere, for automatic acquisition of the CR of senses.

2.2.1 Sense Tagging Based on Conceptual Context Representation and Translations

Now we are ready to propose a heuristic algorithm for tagging bilingual sentences with sense labels. First, the translation morphemes of an MRD definition are added to the CR so that not only the English context, but also the translation (in Chinese for the particular implementation of LDOCE/E-C we will be describing) is considered. For instance, the representation of MONEY-bank contains not only FINANCE words such as *money*, *pay*, *cash*, *capital*, *account*, *charge*, etc., but also the morphemes "銀"

and "行" in the translation of the definition. (See Table 4 for some examples of bilingual context representations for the *bank* senses in LDOCE/E-C.) Subsequently, each CR for a polysemous word is compared with the bilingual sentences. The polysemous word is tagged in favor of the relevant CR that has the most overlap with the bilingual sentences. For instance, consider the case of tagging the instance of *bank* in Example (2) extracted from LDOCE/E-C:

(2) a. the interest in my bank account accrued over the years;
b. 我銀行帳戶的利息逐年有所增加。

Under the assumption of the one-sense-per-translation constraint, the morphemes "銀" and "行" in the translation are sufficient evidence for tagging the instance of *bank* as MONEY-bank. Even if such telling evidence is not present, there nevertheless is a great chance that the sentence contains enough words related to a relevant topic for correct sense tagging to happen. For instance, the FINANCE words, such as *interest* and *account*, in Example (2) lead to the correct sense label MONEY-bank for this instance of *bank*, even when it is not translated as "銀行." The contextual representation derived from the MRD definition also acts as a safety net when the one-translation-per-sense constraint does not hold. For instance, based on the one-translation-per-sense constraint, the instance of *star* in Example (3) can not be labeled as ENTERTAINMENT-star because both the ENTERTAINMENT and HEAVENLY-BODY senses of *star* are translated as "星":

(3) a. she is a star with the theatre company;
b. 她是劇團的紅星。

In such an event, the ENTERTAINMENT words, such as *theatre* and *company*, nonetheless result in the correct sense label: ENTERTAINMENT-star.

A bilingual example in the MRD, or text in a bilingual corpus, can be tagged in the way described above, word by word and sentence by sentence. Unambiguous words with only one sense label are tagged as such. Tagging is done only for content words within the scope of this work. Function words can be treated similarly [Chang, Hsu and Chen 1996]. Sentences in English tagged as training materials can facilitate acquisition of WSD knowledge. The method for tagging a bilingual training corpus is summarized as Algorithm 2. Table 5 shows the result of applying Algorithm 2 to some LDOCE/E-C examples.

Table 4. *Bilingual contextual representations for* bank *senses based on conceptual context representations from definition and dictionary translation.*

Sense Division *S*	Context on $CCR(D_{bank, s})$
MONEY	{銀, 行, money, pay, cash, capital, account, charge ..., pay, bond, bill, charge ..., money, cash, fund, check, ... }
RIVER	{岸, 堤, 沙, 洲, lake, land, river, shore, stream, beach, ..., boat, ship, craft, port, ..., place, edge, road, border, ..., rock, stone, clay, soil, ..., fish, crab, coral, shell, fur, ..., chicken, duct, goose, seabird, ... }
EARTH	{田, 埂, universe, space, planet, constellation, ..., lake, land, river, shore, stream, beach, ..., farming, field, crop, stock, productive, ..., place, edge, road, border, ..., playgroup, school, college, classroom, ... }
PILE	{一, 塊, 一, 團, weather, climate, sky, cloud, fog, steam, ..., roof, chimney, ceiling, wall, door, ground, ..., rock, stone, clay, soil, ..., universe, space, planet, constellation, …, transport, vehicle, car, motorcar, transit, ... }
ROAD	{邊, 坡, transport, vehicle, car, motorcar, transit, ..., place, edge, road, border, ..., lake, land, river, shore, stream, beach, ... }
ROW	{一, 排, transport, vehicle, car, motorcar, transit, … , printing, sign, letter, code, …, sail, caravan, itinerary,…, song, melody, dance, …, road, street, ……... }
MEDICINE	{血, 庫, blood, trunk, breast, back, buttock, waist, ..., patient, examine, diagnose, soothe, ... }
GAMBLE	{莊, 家, athletics, run, jump, ride, game, round, ..., ball game, shoot, golf, pitch, football, ..., cards, pack, suit, heart, club, poker, dice, ... war, warfare, conflict, fight, battleground, ... }

Table 5. *Results of sense tagging.*

Example	The interest in my bank account accrued over the years.	
Translation	我銀行帳戶的利息逐年有所增加。	
Tagged Keywords	interest/**Je**, bank/**Je**, account/**Je**, accrue/**Nd**	
Gloss for Topics	Je	Banking
	Nd	Size

Algorithm 2: Labeling bilingual training corpus

Step 1: Form contextual representation *CR(W, S)* of sense *S* of word *W* with definition *D* and translation *T* as follows:

$$CR(W, S) = LCR(D_{w, s}) + CCR(D_{w, s}) + LCR(T_{w, s}).$$

Step 2: For each word W in an example sentence E, compute the similarity between its context and translation, C_E, and each of the contextual representations $CR(W,S)$ based on the Dice Coefficient:

$$\text{Sim}(C_E, CR(W,S)) = \sum_{c \in C_E} \frac{2 \times \text{In}(C, CR(W,S))}{|C_E| + |CR(W,S)|},$$

where $\text{In}(a, B)$ = the weight of a in B, if $a \in B$ and

0, otherwise.

Step 3: Label W in E with S^* such that $\text{Sim}(C_E, CR(W,S^*))$ is maximized;

$$\text{Sim}(C_E, CR(W,S^*)) = \underset{L}{\text{Max}}\ \text{Sim}(C_E, CR(W,L))$$ and is greater than a certain threshold.

2.2.2 Acquiring Contextual Representations for Example Sentences

Lexicalized and conceptualized CR can be constructed from tagged MRD examples in a fashion similar to that described in Section 2.1 for MRD definitions. Given an ambiguous word W labeled with sense S in a set of example sentences $E_{w,s}$, every content word appearing in E is gathered to form $LCR(E_{w,s})$, shown as follows:

$$LCR(E_{w,s}) = \{ x \mid x \in E_{w,s} \text{ and } x \text{ is not a function word} \}.$$

Table 6 shows some of the contextual words in the LDOCE examples that appear in the context of each of eight *bank* senses. Notice that the entry for MONEY-bank contains many strong collocates, such as (*rob*, *bank*), (*bank*, *account*), etc. These collocates are potentially very helpful for WSD. Although some of the contextual words merely repeat information in the definition-based representation, $LCR(D_{w,s}) + CCR(D_{w,s})$, many do provide new information. For instance, fifteen instances of *river* reaffirm the defining word *river* as an important collocate for RIVER-bank, while contextual words such as *north*, *east*, *deer*, and *vole* provide additional, richer context.

Table 6. *Lexicalized contextual representations for* bank *from the set of LDOCE examples* E.

Sense Label S	Context (with frequency) in $LCR(E_{bank,s})$
MONEY	rob(23), account(15), money(8), criminal(6), interest(5), keep(5), paper(4), police(4), robber(4), thief(4), cheque(3), ...
RIVER	river(15), city(2), north(2), stream(2), east(1), air(1), deer(1), south(1), sea(1), vole(1), ...
EARTH	build(2), earth(2), flood(1), rise(1), water(1), ...
PILE	cloud(2), dark(2), heavy(1), storm(1), ...
ROAD	moss(2), wood(2), rest(1), sit(1), ...
ROW	-
MEDICINE	-
GAMBLE	-

In the previous section, we showed that these contextual words are neither frequent nor necessarily likely to recur. However, when viewed as representing a typical topic or concept, they certainly are recurring. For instance, although there is only one instance of *north bank* in LDOCE examples, there are quite a few *south bank*, and *right bank* instances, all of which signal a recurring context of the DIRECTION concept. Therefore, it is a good idea to derive a conceptualized contextual representation from the set of examples *E* relevant to a sense label *S*. For instance, representing the co-occurring concept of the DIRECTION with RIVER- bank, $CCR(E_{\text{bank, river}})$ would contain such words as *east, west, south, north, left, and right*, etc.:

$$CCR(E_{\text{bank, river}}) = \{ \textit{east, west, south, north, left, right}, \ldots \}.$$

For this purpose, we again turn to the information retrieval (IR) technique. Since the LDOCE in general strictly uses words in the controlled vocabulary for both definitions and examples, the same method described by Chen and Chang [1998b] for forming conceptual characterization of MRD definitions also works for MRD examples. Table 7 shows a list of topical words that characterize the context of each of the eight *bank* senses based on sense tagged LDOCE examples. The results obtained using an IR-based method seem to characterize the context in a general way that can be very useful for WSD.

Table 7. *Conceptualized contextual representation for* bank *from the set of LDOCE examples E.*

Sense Division *S*	Related Topics	Context on $CCR(E_{\text{bank, s}})$
MONEY	Je(0.45), Jf(0.33), Jd(0.22)	{officer, cop, detective, guard, protect, gangster, hoodlum, larceny, hijacking, burglar, steal, fraud, swindle, …}
RIVER	Ld(0.45), Mf(0.26), Me(0.14), Hc(0.07), Af(0.05), Ad(0.04)	{east, west, north, south, up, down, erode, elk, moose, rat, mouse, rabbit, hare, … }
EARTH	La(0.36), Ld(0.24), Eg(0.20), Me(0.12), Ie(0.08)	{tide, ebb, current, spate, …}
PILE	Lc(0.59), Db(0.13), Hc(0.09), La(0.09), Md(0.09)	{fog, steam, haze, dew, mist, …}
ROAD	Md(0.45), Me(0.38), Ld(0.17)	{forest, jungle, hole, crack, …}
ROW	Md(0.49), Gd(0.18), Mc(0.16), Kb(0.12), Me(0.06)	-
MEDICINE	Bd(0.70), Bj(0.30)	-
GAMBLE	Ke(0.35), Kh(0.28), Kf(0.23), Cn(0.14)	-

2.3 Combining Definition-based and Example-based CR

Definition-based and example-based CR as described in Sections 2.1 and 2.2 can be put together to form a combined CR for acquiring word sense. For simplicity, we merge the two to produce the final MRD-based CR. For a polysemous word *W* and a relevant word sense *S*, with the definition *D* of sense

S and the set of examples *E* containing an instance of *S*, the contextual representation *CR*(*W*, *S*) can be represented as follows:

$$WORD(W, S) = LCR(D_{w,s}) + CCR(D_{w,s}) + LCR(E_{w,s}) + CCR(E_{w,s}),$$

where$LCR(D_{w,s})$ is the lexical contextual representation derived from definition *D*,
$CCR(D_{w,s})$ is the conceptual contextual representation derived from definition *D*,
$LCR(E_{w,s})$ is the lexical contextual representation derived from the set of examples *E*,
and $CCR(E_{w,s})$ is the conceptual contextual representation derived from the set of examples *E*.

To take into account the significance of each contextual word in *CR*(*W*, *S*), the IR technique for weighting index terms for relevancy can be applied here to good effect. Using the IR analogy, the collective context of each word sense is viewed as a document, and the relevance of a contextual word *t* to a sense *S* of word *W* depends on its term frequency *tf* and inverse document frequency *idf*. The term frequency *tf* is the number of instances of *t* in *WORD*(*W*, *S*), and *idf* is the percentage of *CR*s in which an instance of *t* appears. The relevancy of a contextual word is estimated using the commonly used scheme: *tf* ×*idf*. Experiments show that the simple scheme tends to give a high weight to strong collocations, such as (*rob*, MONEY-bank) and (*river*, RIVER-bank), thus leading to a representation that is potentially very effective for WSD.

We sum up the above descriptions and outline the procedure as Algorithm 3. The algorithm combines definition-based and example-based CR into an integrated contextual representation *CR*(*W*, *S*) for the sense *S* of the polysemous word *W*.

Algorithm 3: Combining definition-based CR

Step 1: Given a polysemous word *W*, one of its senses *S* and a collection of bilingual examples *C*, run Algorithms 1 and 2 to obtain $LCR(D_{w,s})$, $CCR(D_{w,s})$, $LCR(E_{w,s})$ and $CCR(E_{w,s})$, where *E* is a set of examples that each contain an instance of *S*.

Step 2: Merge the following word list for W and *S*:
$WORD(W, S) = LCR(D_{w,s}) + CCR(D_{w,s}) + LCR(E_{w,s}) + CCR(E_{w,s})$.

Step 3: For each *WORD*(*W, S*), compute a list of distinct words *X* with weight $W_{X,S}$ as follows:
$CR(W, S) = \{ X (W_{X,S}) \mid X$ is a distinct word in $WORD(W, S)\}$,
where $tf_{X,S}$ = the frequency of *X* in *WORD*(*W, S*),
idf_X = 1/the percentage of senses *S* such that $X \in WORD(W, S)$,
$W_{X,S} = tf_{X,S} \times idf_X$.

Step 4: The weights $W_{X,S}$ in *CR*(*W, S*) for each word sense *S* are normalized to a sum of 100.

An Illustrative Example

In the following, we will demonstrate how Algorithm 3 works. Given a MONEY-bank sense, the integrated *CR*(*bank*, MONEY) can be acquired by doing the following (where the numbers in parentheses following collocates denote the frequency):

Step 1: After running Algorithms 1 and 3, we obtain the following:

$LCR(D_{\text{bank.4.n.1}})$ = {place, money, keep, pay, demand, activity },

$CCR(D_{\text{bank.4.n.1}})$ = {money, pay, cash, capital, account, charge, ...
pay, bond, bill, charge, ...
money, cash, fund, check, ... } ,

$LCR(E_{\text{bank.4.n.1}})$ = {rob(23), account(15), money(8), criminal(6), interest(5),
keep(5), paper(4), police(4), robber(4), thief(4),
cheque(3), ...}, and

$CCR(E_{\text{bank.4.n.1}})$ = {officer, cop, detective, guard, protect, gangster, hoodlum,
larceny, hijacking, burglar, steal, fraud, swindle, …}.

Step 2: *WORD*(*bank*, MONEY) =

$$LCR(D_{\text{bank.4.n.1}}) + CCR(D_{\text{bank.4.n.1}}) + LCR(E_{\text{bank.4.n.1}}) + CCR(E_{\text{bank.4.n.1}}).$$

Similar calculations can be performed for other senses of *bank* to obtain *WORD*(*bank*, RIVER), *WORD*(*bank*, EARTH), etc.

Step 3: Compute *tf* and *idf* for each distinct word in *WORD*(*bank, S*). For instance, there are 16 instances of *account* in *WORD*(*bank*, MONEY) and no other word list *WORD*(*bank, S*), S ≠ MONEY, contains *account*. Thus, we have $tf_{\text{account, MONEY}} = 16$ and $idf_{\text{account}} = 8$. Thus, the weight for account in *CR*(*bank*, MONEY) is $W_{\text{account, MONEY}} = 16 * 8 = 128$.

Step 4: The weight $W_{X, S}$ in *CR*(*W, S*) for each word sense S is normalized to a sum of 100. For instance, the total of the weights *CR*(*bank*, MONEY) = 6274.5; therefore, the normalized weight $W_{\text{account, MONEY}} = 2.04$. Table 8 shows more details about the contextual words and normalized weights in *CR*(*bank, S*) for all bank senses S. The ten top-weighted context words from the CRs of the *bank* senses listed in Table 8 seem to be very relevant to each sense and to have strong collocates listed in BBI [Benson, Benson and Ilson 1993]. These weighted context words form the general CCR knowledge for senses of bank. In the next section, we will show that this knowledge is effective for applying WSD to unrestricted text.

Table 8. *Top-ranked words in combined contextual representation based on definitions and examples of* bank *senses.*

Sense *S*	Context(with weights[4]) on *CR*(bank,*S*)
MONEY	rob(6.17), money(2.17), account(2.04), criminal(1.61), interest(1.40), keep(1.39), pay(1.18), police(1.07), robber(1.07), thief(1.07), …
RIVER	river(5.54), leave(2.18), towards(2.18), ship(1.23), city(1.09), dangerous(1.09), deer(1.09), descend(1.09), excavation(1.09), north(0.73), fish(0.56), …
EARTH	build(3.92), vole(3.92), earth(1.42), rise(0.98), flood(0.73), water(0.65), agricultural(0.20), barn(0.20), farm(0.20), garden(0.20), …
PILE	cloud(2.26), dark(1.97), heavy (0.99), storm(0.64), hall(0.36), shower(0.36), atmosphere(0.18), blizzard(0.18), blow(0.06), breeze(0.06), …
ROAD	moss(4.38), sit (2.19), wood (1.14), rest (1.09), gradient (0.18), junction (0.18), subway(0.08), tunnel(0.08), accelerator(0.06), accident(0.06), …
ROW	call(0.19), page(0.19), classical(0.16), compose(0.16), composition(0.16), leader(0.16), caravan(0.12), porter(0.12), bell(0.11), horn(0.11), …
MEDICINE	crutch(0.48), gut(0.48), abdomen(0.34), abdominal(0.34), ankle(0.34), anal(0.34), anus(0.34), aorta(0.34), pendicities(0.34), armpit(0.34), …
GAMBLE	club(0.43), cup(0.32), loser(0.24), win(0.24), defense(0.20), bet(0.17), champion(0.17), competition(0.17), gamble(0.17), games(0.17), …

3. Word Sense Disambiguating Algorithms

Among the recently proposed WSD systems, almost all have the property that the knowledge obtained is fixed when the system completes the training phase. This means that the acquired knowledge can not be enriched during the course of disambiguation. Such fixed knowledge is referred to as static knowledge. We believe that this property limits WSD performance. We propose lifting this limitation by adjusting the initial acquired knowledge to suit the text at hand. Alternatively, such expanded knowledge is referred to as adaptive knowledge. In this section, we will show how to distinguish between senses of text using adaptive disambiguation techniques. First, we will start with disambiguation of polysemous words in easy (trivial) contexts by using the fundamental knowledge previously acquired from MRD. Next, we will expand the acquired knowledge based on these disambiguated contexts. Finally, we will resolve the senses in the remaining contexts, called hard contexts.

3.1 Disambiguating Polysemous Words in Easy Contexts

The proposed WSD method starts with a simple disambiguation step using the topical CR described in a previous section. For instance, to disambiguate the word *bank* in Examples (4) through (6), the content words in its context are extracted, lemmatized and matched against the contextual representation of each of

[4] Weights for all contextual words of a sense are normalized to a sum of 100.

bank's word senses. Each instance of *bank* is given a sense label in favor of a CR most similar to the context in question. A sense label is assigned only when the match is strong enough and the runner-up sense is sufficiently weak. In the following subsections, we will describe how to distinguish between strong and weak signals. For instance, there is enough overlap between the CR for the instance of MONEY-bank in Examples (4) and (6) to warrant a sense label of MONEY-bank for the two instances of *bank*, but the match is not strong enough for the instance in Example (4). We call Examples (4) and (6) easy[5] contexts, while Example (5) is a hard[6] context.

(4) … Participation *loans* are those made jointly by the SBA and *banks* or other private *lending* institutions ...

(5) … individual action by every **nation** in position to help, we must squarely face this titanic challenge …

(6) … from *investment* firms all over the **nation**, all of them wanting a part of *shares* that would be *sold* (185,000 to the public at $12.50 with another 5,000 reserved for Morton Foods employers at $11.50 a *share*) there was even a cable in French from a *bank* in Switzerland that had somehow …

In addition, the contextual words closer to an ambiguous word may have greater influence on the sense of a word. For instance, consider Example (7), where the intended sense of *bank* is MONEY. We observe that there are two salient words, *mortgage* and *river*, around an ambiguous word *bank*. The word *mortgage* favors a MONEY sense, while the word *river* favors a RIVER sense. Intuitively, the MONEY sense should be given more favorable consideration since *mortgage* is nearer to the ambiguous word than *river* is. There are various representations for distance-based weights. Here we adopt the metric proposed by Hawking and Thistlewaite [1995] to weigh the relevance of salient words in a text.

(7) … and an effort to get this religious center out of its rut of wild worship into a modern church organization. He emphasized to the Presiding Elder the plan of giving up the old church and moving across the river. The Presiding Elder was sure that that would be impossible. But he told Wilson to "go ahead and try". And Wilson tried. It did seem impossible. The *bank* which held the **mortgage** on the old church declared that the interest was considerably in arrears, and the real estate people said flatly that the land across the **river** was being held for an eventual development for white working people who were coming in, and that none would be sold to colored folk. When it was proposed to rebuild the church, Wilson found that the terms for …

To sum up, we outline a general WSD method using MRD-based contextual representation as Algorithm 4 for labeling an instance of a polysemous word *W* in a particular context *CON*(*W*).

---

[5] The algorithm for identifying easy contexts is Algorithm 4.

[6] The algorithm for resolving hard contexts is Algorithm 5.

Algorithm 4: (*StaticSense*) WSD using MRD-based contextual representation

Step 1: Preprocess the context and produce a list of lemmatized content words *CON*(*W*) in *W*'s context.

Step 2: For each sense S of W, compute the similarity between the context representation *CR*(*W, S*) and topical context *CON*(*W*).

$$\text{Sim}\,(CR(W,S), CON(W)) = \frac{\sum_{t \in M} (W_{t,s} + W_t)}{\sum_{t \in CR(W,S)} W_{t,s} + \sum_{t \in CON(W)} W_t},$$

where $M = CR(W,S) \cap CON(W)$,

$W_{t,s}$ = the weight of a contextual word t with sense S in *CR*(*W*),

W_t = the weight of t in $CON(W) = \frac{1}{\sqrt{|X_t|}}$,

X_t = the distance from *t* to *W* in number of words,

$$S^*(W, CON(W)) = \arg\max_s \text{Sim}\,(CR(W,S), CON(W)),$$

$$S''(W, CON(W)) = \arg\max_s \{\text{Sim}\,(CR(W,S), CON(W)) \mid \text{Sim}\,(CR(W,S), CON(W)) < S^*(W, CON(W))\},$$

$$TSCORE(W, CON(W)) = \frac{S^*(W, CON(W))}{S''(W, CON(W))},$$

RANK-S(*W*, *CON*(*W*)) = the rank of *S**(*W*, *CON*(*W*)) among all *S**(*X*, *CON*(*X*)) for all n instances of polysemous word X and context *CON*(*X*),

RANK-T(*W*, *CON*(*W*)) = the rank of TSCORE(*W*, *CON*(*W*)) among TSCORE(*X*, *CON*(*X*)) for all n instances and context of polysemous word *X*.

Step 3: Construct the set of the triples T, where
T = { (*W*, *S*, *CON*(*W*)) | *S* = *S**(*W*, *CON*(*W*)) such that
RANK-S(*W*, *CON*(*W*)) ≤ *n/c* and
RANK-T(*W*, *CON*(*W*)) ≤ *n/c*,
where the constant $c \geq 1$ }[7].

Step 4: *DEFAULT*(*W*)= *S* such that the count of (*W, S, CON*(*W*)) ε *T* is the largest among all the senses of *W*.

Step 5: Assign (*W*, *CON(W)*) as the relevant sense *S* if (*W, S*, *CON(W)*) is in *T*, and assign *DEFAULT(W)* otherwise.

3.2 Adapting the Knowledge Base to Fit the Text

The adaptive approach to WSD hinges on two assumptions. First, we assume that it is possible to build

[7] We use the WSD results of the top-ranking *c*'th instances in S* as well as TSCORE values which are more reliable. For instance, setting *c* to 2 amounts to taking the top-ranking 25% quantile of the test cases.

an initial general knowledge base so that a substantial portion of disambiguated text can be used to adapt the knowledge base to fit the text itself. The second condition for the adaptive approach to be feasible is that there is indeed new and effective information to be gained from the partially disambiguated text. In this section, we will first show the kinds of contexts in the Brown corpus and WSJ articles in which word sense ambiguity can be confidently resolved by using an MRD-based knowledge base. In these contexts, one will find an abundance of rich task-specific information not easily covered in a general or static knowledge base. We will also justify the use of contextual information and a task-specific default for WSD.

3.2.1 Discovering Task-specific Contextual Information

There is indeed an abundance of new and useful contextual information for word sense to be gained from typical, easy contexts. Such information can be extracted as long as ambiguity in these typical contexts can be interpreted successfully. For instance, the Brown corpus passage reproduced here as Example (8) is obviously very typical of MONEY-bank with salient words such as *accounts*, *stocks* and *property* in its context. Without a doubt, this instance of MONEY-bank can be resolved successfully using the kind of MRD-based knowledge base described in Section 3.1. Even though the overall context of this instance of MONEY-bank is a general one, it nevertheless contains many words, such as *law* and *state*, not in the MRD-based knowledge base. Such words might very well be incidental and have no intrinsic relation with the sense. For instance, the word *law* might just as likely be associated with RIVER-bank as MONEY-bank. Without much stretching of the imagination, it is possible to think of a likely event where *the state of Texas passes a law to declare an outer bank off limits to commercial development*. However, more often than not, these unexpected words will indicate real recurring contexts of word sense, either generally or in a task-specific way. Therefore, adapting the knowledge base to fit such a context is beneficial for WSD. For instance, the instances of *tree* and *camping* in the context of RIVER-bank in Example (9) seem to be reasonable additions to *CR*(*bank*, RIVER) in the sense that *tree* and *camping* are, in general, more strongly associated with RIVER-bank than with MONEY-bank. Even if that assertion generally does not hold, adding *tree* and *camping* to *CR*(*bank*, RIVER) as a way of adapting the knowledge base is still beneficial since it is likely to be valid in the very text where this association is discovered. The same argument holds for the local cue of *through* in the context of PILE-bank in Example (10), and for the instances of *donor* and *transfusion* in the context of MEDICINE-bank in Example (11). (See Table 9 for further details.)

(8) … 63 million dollars at the end of the current fiscal year next Aug. 31. He told the committee the measure would merely provide means of enforcing the escheat law which has been on the books "since Texas was a republic". It permits the state to take over *bank* **accounts**, **stocks** and other personal **property** of persons missing for seven years or more. The bill, which Daniel said he drafted personally, would force banks, insurance firms, pipeline companies and other …

(9) … On shooting preserves? Ask Sammy Shooter. WE WERE CAMPING a few weeks ago on Cape Hatteras Campground in that land of pirates, **seagulls** and **bluefish** on North Carolina's famed outer *bank*s. This **beach** campground with no *trees* or hills presents a constant **camping** show with all manner of equipment in actual use. With the whole camp exposed to view we could see the variety of canvas shelters in which Americans are camping now. There were …

(10) … to let down through the overcast and see the ground before it hit him. Bob Fogg didn't have today's advantages of Instrument Flight and Ground Control Approach systems. At the end of the calculated time he'd nose the Waco down **through** the **cloud** *bank* and hope to break through where some feature of the winter landscape would be recognizable. Usually back in Concord by noon, there was just time to get partially thawed out, refuel, and grab a bit of Mrs. Fogg's …

(11) … agreed, but explained that it would be necessary first to check Fred's blood to ascertain whether or not it was of the same type as Papa's. To give a **patient** the wrong type of blood, said the **doctor**, would likely kill him. That was in the days before **blood** *bank*s, of course, and **transfusions** had to be given directly from *donor* to **patient**. One had to find a donor, and usually very quickly, whose blood corresponded with the patient's. And then it took considerably

Table 9. *Samples of disambiguated topical contexts of* bank *in the Brown corpus.*

Sense	Example No.	Typical, Easy Context of Various Senses of *bank*	General Topical Context	Task-specific Context
MONEY	(9)	… It permits the state to take over *bank* **accounts**, **stocks** and other personal **property** of persons missing for seven years or more. …	Account Stock Property	law bill
RIVER	(10)	… WE WERE CAMPING a few weeks ago on Cape Hatteras Campground in that land of pirates, **seagulls** and **bluefish** on North Carolina's famed outer *bank*s. …	Seagull Bluefish Hill	tree camping
PILE	(11)	… At the end of the calculated time he'd nose the Waco down *through* the **cloud** *bank* and hope to breakthrough where some feature of the winter landscape would be recognizable. …	Cloud	flight through
MEDICINE	(12)	… That was in the days before **blood** *bank*s, of course, and *transfusions* had to be given directly from *donor* to **patient**. …	blood doctor patient	donor transfusion

3.2.2 Using the Default Sense

The distribution of senses of a word might not follow Zipf's law because their rank-frequency plot does not follow the power-law well, and it is often quite skewed even in a balanced corpus. In the Brown corpus,

60% of the instances of twelve polysemous words are the top-ranking sense of the word, according to an experimental report by Luk [1995].

Generally, the top-ranking sense of a word is corpus-dependent. Table 10 presents some statistics about the distribution of senses in different corpora. For instance, we find that CURIOSITY-interest is favored over MONEY-interest 194 to 49 in the Brown corpus, while preference is reversed with counts of 53 and 122 in the WSJ corpus. On the other case, GRAMMAR-sentence is favored over JUDGEMENT-sentence 22 and 10 in the Brown corpus while preference is reversed with counts of 1 to 11 in the WSJ corpus. Using a fixed default would be disastrous for *interest* or *sentence* in at least one of these corpora. The adaptive method alternatively uses a set of disambiguated samples from the text in question to estimate the default.

Table 10. *Skewed sense distribution is corpus dependent*

Word	Sense	Brown	WSJ
Interest	MONEY	49	122
	CURIOSITY	194	53
Sentence	GAMMAR	22	1
	JUDGEMENT	10	11
Bass	MUSIC	15	2
	FISH	1	0

3.3 The Adaptive WSD Algorithm

We are now ready to present a new adaptive approach to WSD based on the fundamental knowledge base acquired from MRD. Previous sections have already shown how such a knowledge base can be built and described its advantages. We will show one way of using a MRD-based knowledge base for WSD. Although the knowledge base does not guarantee high precision and 100% coverage, a substantial portion, say 50%, can be disambiguated at a high precision rate. In this section, we will show how such a level of coverage and high precision can be put to use in an adaptive way to maintain the same high precision rate at 100% coverage. We will first describe the adaptive algorithm. Examples will be given in Section 3.4 to illustrate how the algorithm works and to give some idea of the potential effectiveness of adaptation.

The algorithm starts with an initial disambiguation step using the knowledge base derived from the MRD. An adaptation step follows which produces a knowledge base from the partially disambiguated text. Finally, the undisambiguated part is disambiguated according to the adapted knowledge base. Algorithm 5 gives a formal and detailed description.

Algorithm 5: (*AdaptSense*) Adaptive WSD

Step 1: Run Algorithm 4 to obtain triples T_1 of word, word sense and context.

Step 2: From the selected triples $(W, S, CON(W)) \in T_1$, compute a new set of contextual representations:

$WORD(W,S) = \{ u \mid u \in CON(W) \text{and } (W, S, CON(W)) \in T_1 \}$.

Step 3: Build the contextual representation $CR(W,S)$ of sense S of word W from $WORD(W, S)$ according to Algorithm 3.

$DEFAULT(W) = S$ such that the count of $(W, S, CON(W)) \in T_1$ is the highest among all the senses of W.

Step 4: For all the instances of polysemous W and its $CON(W)$ such that $(W, S, CON(W))$ is not in T_1 (for all senses S of W),

$$\text{Sim}(CR(W,S), CON(W)) = \frac{\sum_{t \text{ in } M} (W_{t,s} + W_t)}{\sum_{t \text{ in } CR(W,S)} W_{t,s} + \sum_{t \text{ in } CON(W)} W_t},$$

where $M = CR(W,S) \cap CON(W)$,

$W_{t,s}$ = the weight of a contextual word t with sense s in $CR(W,S)$,

W_t = the weight of t in $CON(W) = \frac{1}{\sqrt{|X_t|}}$,

X_t = the distance from t to W in number of words.

$S^*(W, CON(W)) = \arg\max_s \text{Sim}(CR(W,S), CON(W))$,

$S''(W, CON(W)) = \arg\max_s \{ \text{Sim}(CR(W,S), CON(W)) \mid$

$\text{Sim}(CR(W,S), CON(W)) < S^*(W, CON(W)) \}$,

$$TSCORE(W, CON(W)) = \frac{S^*(W, CON(W))}{S''(W, CON(W))},$$

$RANK\text{-}S(W, CON(W)) =$ the rank of $S^*(W, CON(W))$ among all $S^*(X, CON(X))$ for all n instances of polysemous word X and context $CON(X)$,

$RANK\text{-}T(W, CON(W)) =$ the rank of TSCORE$(W, CON(W))$ among TSCORE$(X, CON(X))$ for all n instances and context of polysemous word X.

Step 5: Construct the set of triples T_2, where

$T_2 = \{ (W, S, CON(W)) \mid S = S^*(W, CON(W))$ such that $RANK\text{-}S(W, CON(W)) \le n/c$ and RANK-T$(W, CON(W)) \le n/c$, where the constant $c \ge 1 \}$.

Step 6: Assign (W, $CON(W)$) to the relevant sense S, such that

$(W, S, CON(W)) \in T_1$, or

$(W, S, CON(W)) \in T_2$, or

$DEFAULT(W)$, otherwise.

3.4 An Illustrative Example

To show how Algorithm 5 works in an adaptive fashion, we will consider the case of disambiguating the Brown corpus, focusing on the polysemous word *bank*. For this purpose, we will describe step by step how the algorithm operates on the two following passages in the Brown corpus containing an instance of *bank*. The two passages are reproduced here as Examples (12) and (13), showing a context window of 50 words before and after the polysemous word which is used in the algorithm for disambiguation.

(12) … to face charges of assault and robbery, Portland detectives said Friday. Mrs. Lavaughn Huntley is accused of driving the getaway car used in a robbery of the Woodyard Bros' Grocery, 2825 E. Burnside St., in April of 1959. Her husband, who was sentenced to 15 years in the federal prison at McNeil Island last April for robbery of the hillsdale branch of Multnomah *Bank*, also was charged with the store holdup. Secret Grand Jury indictments were returned against the pair last week, Detective Murray Logan reported. The Phoenix arrest culminates more than a year's investigation by Detective William Taylor and other officers. Taylor said Mrs. Huntley and her husband also will be questioned about …

(13) … Of cattle in a pasture without throwin' 'em together for the purpose was called a "pasture count". The counters rode through the pasture countin' each bunch of grazin' cattle, and drifted it back so that it didn't get mixed with the uncounted cattle ahead. This method of countin' was usually done at the request, and in the presence, of a representative of the *bank* that held the papers against the herd. The notes and mortgages were spoken of as "cattle paper". A "book count" was the sellin' of cattle by the books, commonly resorted to in the early days, sometimes much to the profit of the seller. This led to the famous sayin' in the Northwest of the "books won't freeze". This became a common byword durin' the …

Step1: Identifying an easy context

This step corresponds to five substeps of Algorithm 4. First, only the salient words that are in $CR(bank, S)$ for S in {MONEY, RIVER, EARTH, PILE, ROW, ROAD, MEDICINE, GAMBLE} are of interest; all other words are thrown out for now. To calculate similarity values, the weights for these words with respect to relevant senses are pulled out from the initial knowledge base. Tables 11 (a) and (b) show these words, their position relative to *bank*, and their weights according to a knowledge base extracted from LDOCE.

Table 11(a). *Weights for salient words in Example* (*12*) *for* bank *in the initial WSD stage.*

	X_t	W_t	$W_{t,S}$ in CR (bank, S)							
			S_{MONEY}	S_{RIVER}	S_{EARTH}	S_{PILE}	S_{ROW}	S_{ROAD}	$S_{MEDICINE}$	S_{GAMBLE}
drive	-46	0.15	-	0.81	0.02	0.02	0.07	0.01	-	0.04
getaway	-44	0.15	1.37	-	-	-	-	-	-	-
car	-43	0.15	0.94	-	-	0.81	0.12	0.07	-	-
robbery	-39	0.16	0.81	-	-	-	-	-	-	-
husband	-24	0.20	1.07	-	-	-	-	-	-	-
year	-18	0.24	0.81	-	-	-	-	-	-	-
prison	-14	0.27	2.04	-	-	-	-	-	-	-
robbery	-7	0.38	2.31	-	-	-	-	-	-	-
branch	-3	0.58	3.81	-	-	-	-	-	-	-
bank	0									
charge	3	0.58	3.58	-	-	-	-	-	-	-
return	13	0.28	1.77	-	-	-	-	-	-	-
week	18	0.24	1.34	-	-	-	-	-	-	-
report	22	0.21	-	-	-	-	-	0.90	-	-
year	30	0.18	0.81	-	-	-	-	-	-	-
husband	45	0.15	1.07	-	-	-	-	-	-	-

The context of Example (12) resembles the CR of MONEY-bank the most. Table 11(a) indicates clearly that very salient words in *CR*(*bank*, MONEY-bank), such as *robbery*, *branch*, and *charge*, occur in close proximity to the word *bank*. Although words related to other senses, such as *drive* and *report*, do occur, they are fewer and are located at quite a longer distance. It is not surprising that the similarity of this context with MONEY-bank and the *t*-score ranks high enough for this instance to be included in T_1.

On the other hand, Example (13) does not resemble the CR of any particular sense of *bank* more than it does those of other senses. That is evident from the weights shown in Table 11(b). The only indicative word *representative* is not enough to enable interpretation of the intended sense of MONEY-bank. All other words are either not in any CRs or ambivalent (*hold*, *note* and *paper*), indicating a number of senses competing with MONEY-bank. Hence, the similarity of this context with MONEY-bank and the *t*-score do not rank high enough for this instance to be included in T_1.

Table 11(b). *Weights for salient words in Example* (*13*) *for* bank *in the initial WSD stage.*

Word	X_t	W_t	$W_{t,s}$ in CR (bank, S)							
			S_{MONEY}	S_{RIVER}	S_{EARTH}	S_{PILE}	S_{ROW}	S_{ROAD}	$S_{MEDICINE}$	S_{GAMBLE}
call	-50	0.14	-	-	-	-	-	0.99	-	-
pasture	-48	0.14	-	-	1.00	-	-	-	-	-
count	-47	0.15	0.82	-	-	-	-	-	-	0.86
counter	-45	0.15	-	-	-	-	-	-	-	0.92
ride	-44	0.15	-	-	-	-	-	-	-	0.97
pasture	-41	0.16	-	-	1.00	-	-	-	-	-
ahead	-20	0.22	-	0.94	-	-	-	-	-	-
representative	-3	0.58	3.04	-	-	-	-	-	-	-
bank	0									
hold	2	0.71	-	-	3.65	-	-	3.03	-	3.04
paper	4	0.50	1.07	0.81	-	0.81	-	0.82	-	-
note	9	0.33	1.79	-	-	-	-	1.58	-	-
paper	17	0.24	1.07	0.81	-	0.81	-	0.82	-	-
book	19	0.23	0.93	-	-	-	-	0.89	-	-
count	20	0.22	0.82	-	-	-	-		-	0.86
book	28	0.19	0.93	-	-	-	-	0.89	-	-
profit	40	0.16	0.84	-	-	-	-		-	-
lead	45	0.15	-	0.81	-	0.82	0.92	0.89	-	-

Step 2: Computing a new contextual representation set based on an easy context

With the instance *bank* in Example (12) resolved to MONEY-bank, the following triple is created and added to T_1.

> (*bank*, MONEY-bank, "to face charges of assault and robbery, Portland detectives said Friday. Mrs. Lavaughn Huntley is accused of driving the getaway car used in a robbery of the Woodyard Bros' Grocery, 2825 E. Burnside St., in April of 1959. Her husband, who was sentenced to 15 years in the federal prison at McNeil Island last april for robbery of the Hillsdale branch of Multnomah **Bank**, also was charged with the store holdup. Secret Grand Jury indictments were returned against the pair last week, Detective Murray Logan reported. The Phoenix arrest culminates more than a year's investigation by Detective William Taylor and other officers. Taylor said Mrs. Huntley and her husband also will be questioned about")

From the triples T_1, a list *WORD*(*S*) of contextual words for each sense *S* of word *bank* and the most frequent sense *DEFAULT*(*bank*) are calculated. Therefore, the contextual words in the triple from Example (12) will be lemmatized. With stop words removed, we obtain a list like the following:

WORD(MONEY-bank) = {face, charge, assault, robbery, portland, detectives, say, Friday, mrs, lavaughn, huntley, accuse, drive, getaway, car, use, robbery, woodyard, bros, grocery, burnside, st, april, husband, sentence, year, federal, prison, mcneil, island, last, april, robbery, hillsdale, branch, multnomah, charge, store, holdup, secret, grand, jury, indictment, return, against, pair, last, week, detective, murray, logan, report, phoenix, arrest, culminate, year, investigation, detective, william, taylor, officer, taylor, say, mrs, huntley, husband, question, … }

Step 3: Assigning weight to the contextual representation

From the word lists for all senses of *bank*, the new set of CRs can be derived. The *CR*(*bank*, S) for the word sense *S* basically consists of every word in *WORD*(*S*) associated with a weight. Weights are assigned in favor of contextual words frequently occurring in the context of a particular word sense and that of a smaller number of other senses. For instance, the word *cooperative* occurs very frequently and only in the context of MONEY-bank in the part of the Brown corpus resolved in Step 1. Table 12 (b) shows a completely new set of CRs for *bank*'s senses which are obviously quite different from the LDOCE-based CR shown in Table 12(a). Intuitively, this set of CRs should be more relevant to the Brown corpus than the MRD-based ones.

According to our experiment, there are quite a number of *bank* instances in the Brown corpus that are very typical and can be reliably resolved using LDOCE-based contextual representation. Those instances are predominately resolved as MONEY-bank. Therefore, we have *DEFAULT*(*bank*) = MONEY-bank.

***Table 12* (*a*).** *Selected sample of initial knowledge for* bank *senses.*

Sense	Top-ranking Contextual Words (with weights)
MONEY	rob(6.17), money(2.17), account(2.04), interest(1.40), pay(1.17), robber(1.07), month(0.80), robbery(0.81), prison(0.81), year(0.81), charge(0.58), …
RIVER	river(5.54), ship(1.23), deer(1.09), hunter(1.09), drive(0.81), fish(0.55), air(0.54), hill(0.44), east(0.36), south(0.36), boat(0.13), boatman(0.13), …
EARTH	build(3.91), water(0.65), sky(0.29), plant(0.19), west(0.17), north(0.11), south(0.11), sidewalk(0.02), street(0.02), drive(0.02) , bridge(0.01), …
PILE	wet(0.29), basin(0.06), window(0.06), table(0.03), clay(0.02), cloth(0.02), drive(0.02), …
ROAD	moss(4.38), sit(2.19), wood(1.14), subway(0.18), car(0.07), drive(0.01), …
ROW	car(0.12), letter(0.09), write(0.09), visit(0.08), drive(0.07), column(0.04), story(0.04), …
MEDICINE	blood(0.33), body(0.33), shoulder(0.33), patient(0.14), doctor(0.07), course(0.03), …
GAMBLE	box(0.11), pocket(0.11), play(0.08), check(0.05), point(0.05), drive(0.04), …

Steps 4-6: Disambiguating a hard context

Armed with the new CRs, the instances that do not pass the test in Step 1 are re-evaluated again in Step 4. The similarity for each of those instances, including Example (13), is re-calculated for all possible word senses. The new weights for contextual words in Example (13) are shown in Table 13.

Contrary to the situation in Step 1, where the MRD-based CR is used, there are now more words in the context that are indicative of the intended sense. From the perspective of the new CRs, the words *method*, *usual*, *request*, *paper*, *note*, and *book* all point to the sense of MONEY-bank and not to any other sense. These words either do not exist or ambivalent with respect to the MRD-based CR. As a whole, these words provide enough evidence to reverse the previous inconclusive situation leading to the expected sense of MONEY-bank. In the event that the maximal similarity is lower than a threshold value, the default sense of MONEY-bank is used. In this particular case, the default happens to be correct.

***Table 12* (*b*).** *Selected sample of adaptive knowledge for* bank *senses.*

Sense	Selected Contextual Words (with weights)
MONEY	cooperatives(1.50), department(1.37), affairs(1.07), export-import(0.69), federal(0.60), government(0.54), short-term(0.43), cooperative(0.41), administration(0.38), firm(0.35), sponsor(0.35), …
RIVER	church(0.80), soldier(0.68), dill(0.66), camping(0.54), fame(0.52), outer(0.52), rhine(0.52), motel(0.40), sight(0.40), tree(0.40), camp(0.33), …
EARTH	manchester(4.35), company(3.77), telegraph(3.77), goodwin(3.11), power(1.88), light(1.69), door(1.64), cemetery(1.31), commercial(1.31), dwelling(1.31), electric(1.31), business(1.23), construction(1.23), …
PILE	tiber(3.69), fold(2.83), moonlight(1.84), thick(1.84), anatomy(0.98), bedside(0.98), buckle(0.98), damn(0.98), dancer(0.98), dark(0.58), …
ROAD	-
ROW	feel(8.51), error(5.37), correct(3.58), shareholder(3.47), people(3.36), data(0.89), fund(0.89), funds(0.89), …
MEDICINE	stumbled(4.51), transfusions(3.46), donor(2.25), frail(2.25), child(1.20), laboratory(1.20), neck(1.20), night(1.20), sample(1.20), …
GAMBLE	fraud(4.10), drink(2.85), grade(2.67), stare(2.67), chief(1.42), collusion(1.42), conclusive(1.42), death(1.42), …

4. Experiment and Evaluation

4.1 Experiment

The experimental setup can be described in a number of steps as follows. (1) A set of 13 polysemous words was selected as the target for disambiguation and evaluation. (2) For each of the polysemous

words, a sense division was established based on the LDOCE treatment of relevant nominal senses. The LDOCE's sense division was used largely as is, with only a couple of closely related senses merged. (3) Two sets of text from corpora were gathered as the *test sets*. (4) Two human judges were asked to assign a sense label to each nominal instance of these 13 words in the two test sets. (5) Two WSD programs were written to disambiguate nominal instances of these polysemous words in the test sets. (6) The results of running the two programs on both test sets were compared against those of human assessors. The number of test instances and that of correctly disambiguated ones in these four experiments were tallied to produce a precision rate for each experiment. In the following, we describe each step in turn.

Table 13. *Weights for salient words in Example* (*13*) *after adaptation.*

Word	X_t	W_t	$W_{t,s}$ in CR(bank, S)							
			S_{MONEY}	S_{RIVER}	S_{EARTH}	S_{PILE}	S_{ROW}	S_{ROAD}	$S_{MEDICINE}$	S_{GAMBLE}
cattle	-21	0.22	0.83	1.00	-	-	-	-	-	-
ahead	-20	0.22	-	1.06	-	-	-	-	-	-
method	-18	0.24	0.86	-	-	-	-	-	-	-
usual	-14	0.27	1.53	-	-	-	-	-	2.10	-
request	-10	0.32	1.62	-	-	-	-	-	-	-
representative	-3	0.58	3.06	-	-	-	-	-	-	-
bank	0									
hold	2	0.71	3.01	3.31	-	3.33	-	-	-	-
paper	4	0.50	0.95		-	-	-	-	-	-
herd	7	0.38	-	1.76	-	-	-	-	-	-
note	9	0.33	1.59		-	-	-	-	-	-
mortgage	11	0.30	1.54	1.63	-	-	-	-	-	-
speak	13	0.28	-	1.64	-	-	-	-	-	-
cattle	16	0.25	0.83	1.00	-	-	-	-	-	-
paper	17	0.24	0.95	-	-	-	-	-	-	-
book	19	0.23	0.89	-	-	-	-	-	-	-
cattle	25	0.20	0.83	1.00	-	-	-	-	-	-

(1) Test words

We limited our experiment and evaluation to a set of thirteen words with higher than usual ambiguity. That is due mainly to the fact that the process of evaluation is a difficult and expensive one. It is often difficult to pin down the number of senses allowed for a word in the experiment. For the purpose of comparing results with other approaches, we stick to words that have been studied in various experiments reported in the literature on computational linguistics. These words include *bank*, *bass*, *bow*, *cone*, *duty*, *galley*, *interest*, *issue*, *mole*, *sentence*, *slug*, *star*, and *taste*.

(2) Sense division

The sense division for each of these test words was very crucial in the WSD experiment. We used a sense division based on LDOCE's treatment of the nominal senses of these words. The division is somehow more fine-grained than those used in other WSD studies. This level of sense division is very close to the kind of granularity required for machine translation. For most cases, a word sense has a unique Chinese translation.

(3) Test corpora

We aimed to determine the effectiveness of the proposed approach for unrestricted text and to find out how domain and genre affect WSD. Therefore, we used the Brown corpus and a collection of WSJ articles from October 30 to November 2, 1989 as the test sets. Passages of 100 words centered at an instance of the test words in the two corpora were extracted using a SED program. It is in general not hard to write a regular expression in the SED program to exclude verbal instances, so only a small number of verb cases were extracted. These verbal instances were excluded from the experiment according to the marks made by human judges. For these thirteen words under investigation, we had 846 and 903 passages of nominal senses from the Brown corpus and WSJ test sets, respectively.

(4) Judgement

To be as subjective as possible, we asked two human judges to assign a sense label to each nominal instance of these thirteen words in the two test sets. There were also cases which fell out of the scope of our sense division. Most of these cases used proper nouns, so they bore none of the meaning represented in our sense division. Cases judged to be verbal uses or proper names were removed from the test cases. For instance, the word *bow* in a Brown corpus passage, reproduced here as Example (14), was an instance of a proper name and, therefore, was excluded from the test cases.

(14) … The announcement that the secrets of the Dreadnought had been stolen was made in *Bow* St. police court here at the end of a three day hearing …

(5) Static vs. Adaptive WSD

In the previous sections, we argued in favor of using an MRD-derived knowledge base because we believe that the fundamental information in an MRD can be very helpful for WSD. Despite our belief in the effectiveness of the MRD-derived knowledge base, we also expected that adaptation could improve its effectiveness a bit further. Therefore, we implemented programs for both Algorithms 4 and 5. These two programs were executed in order to disambiguate the test cases in the Brown and WSJ corpora.

4.2 Evaluation

The results of running the two programs on both test sets were compared against those of human assessors. The number of test instances and that of correct assignments in these four experiments were tallied to

calculate the precision rate for each experiment. All results were based on 100% applicability[8]. Statistics for the experimental results are summarized in Tables 14 and 15. Several observations can be made based on the results. First, evidently, the MRD-based knowledge base was reasonably helpful for WSD. The results shown in Tables 14 and 15 indicate that without adaptation, the knowledge extracted from LDOCE and LLOCE could be used to deliver a precision rate of 65.2% for the Brown corpus and 76.6% for the WSJ articles. Second, adaptation indeed helped boost the precision rate by over 5% for the Brown corpus. As for the WSJ test set (see Table 15), adaptation only marginally increased the average precision rate. Closer examination of the results for this test set shows that three words, *bank*, *interest*, and *issue*, dominated the experiment and evaluation results. The precision rate for *bank* was over 95%, which left adaptation with very little room for improvement. The other two words, *interest* and *issue*, were very general and difficult to disambiguate.

5. Discussion

Although it is often difficult to compare results from experiments based on different domains, genres and setups, the experimental results presented here seem to compare favorably with the experimental results reported in previous WSD research. Our adaptive approach could disambiguate with an average precision rate of 71.2% for these thirteen words in Brown and of 76.5% for these words in WSJ. For the Brown corpus, Luk [1995] experimented with the same words we used except for the word *bank* and reported that there were totally 616 instances of these words (slightly less than the 749 instances we found). The precision rate for all instances was 60%. Leacock, Towell and Voorhees [1993] reported a precision rate of 76% for disambiguating the word *line* in a sample of WSJ articles.

[8]Applicability (coverage) denotes the proportion of cases in which the WSD model performed disambiguation.

Table 14. *Disambiguation results for thirteen ambiguous words in the Brown corpus.*

Word	Sample Sizes	StaticSense	AdaptSense	
		# of correct	# of correct in 1st run	# of correct in 2nd run
bank	97	68	71	71
bass	16	16	16	16
bow	12	3	3	2
cone	14	14	14	14
duty	75	67	69	69
galley	4	4	4	4
interest	346	213	228	226
issue	141	67	88	97
mole	4	2	2	2
sentence	32	30	30	30
slug	8	4	6	6
star	46	28	29	29
taste	51	36	36	36
Total	846	552	596	602
precision		65.2%	70.5%	71.2%

Besides the precision rate, a number of interesting features of this approach are also important. First, the proposed disambiguation system is robust and portable, since absolutely no corpus-specific knowledge is needed in the disambiguation procedure. It can be applied readily to test data in a variety of domains and genres with performance rivaling that of methods requiring a substantial training corpus. Second, the proposed approach is considerably more time efficient when compared to other learning strategies. Although the bootstrap approach proposed by Yarowsky [1995] has an element of adaptation to it, his method still requires a long training process to derive a static knowledge base for WSD. The differences between our method and his lie in the initial knowledge, the level of abstraction, and the learning cycle. We propose to exploit rich conceptualized knowledge from MRD at the outset, while the bootstrap method uses merely a couple of word collocations for each sense to start the learning process. Since the bootstrap method aims to derive a word-based conceptual representation with a large parameter space, a very large training corpus is required. The thesaurus used in the proposed approach provides an appropriate level of abstraction and, thus, alleviates the need for a very large corpus. The time required for learning in the two approaches is also quite different. The adaptive approach requires a single round of adaptation for effective WSD, while the bootstrap method needs many rounds of learning.

Table 15. *Disambiguation results for thirteen ambiguous words in the WSJ test set.*

Word	Sample Sizes	StaticSense	AdaptSense	
		# of correct	# of correct in 1st run	# of correct in 2nd run
Bank	370	350	353	353
Bass	2	2	2	2
Bow	-	-	-	-
Cone	-	-	-	-
duty	25	19	22	22
galley	-	-	-	-
interest	221	123	127	122
issue	260	181	177	175
mole	-	-	-	-
sentence	12	11	12	12
slug	-	-	-	-
star	7	3	2	2
taste	6	3	3	3
Total	903	692	698	691
precision		76.6%	77.3%	76.5%

Speedy adaptation is the consequence of using rich conceptualized knowledge to start the learning process. To show that this is truly the case, we have revised Algorithm 5 by adding a second and a third adaptation step and by applying the new CR to a reserved batch of low-ranking instances instead of using defaults. The results obtained using more adaptation steps are shown in Figure 2. The precision rates show that the additional adaptation steps have only a marginal effect.

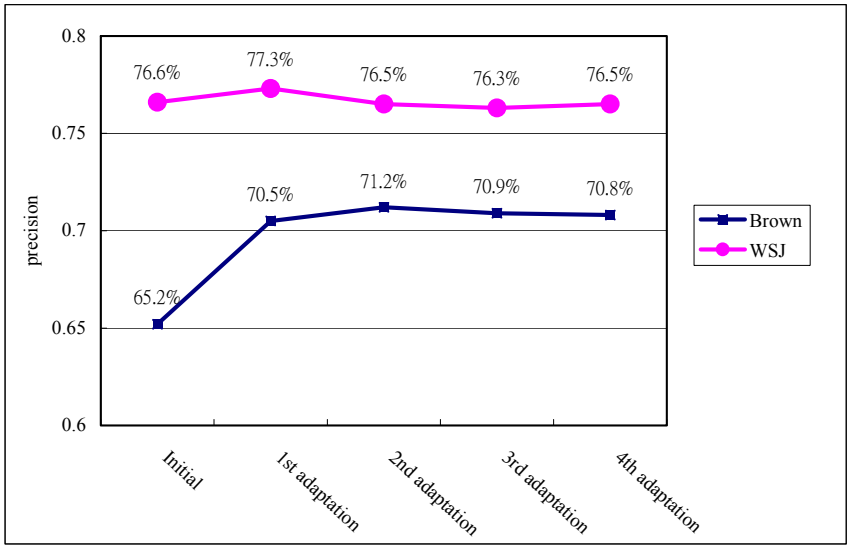


Figure 2 *Average precision rates with and without adaptation.*

One of the limiting factors of this approach is the quality of sense definition in the MRD. Short and vague definitions tend to lead to inclusion of inappropriate topics in the contextual representation. With such inferior CRs, it is not possible to produce enough precise samples in the initial step for subsequent adaptation. For instance, it is difficult to derive appropriate contextual knowledge for the LDOCE senses in (15) since their definitions mainly consist of either function words or very common words:

(15)**interest.1.n.1** a readiness to give attention
issue.1.n.1 the act of coming out
issue.1.n.2 an example of this
issue.1.n.3 something which comes or is given out
issue.1.n.4 the act of bringing out something in a new form

The experiment and evaluation results show that adaptation is most effective when a high-frequency word with contrasting senses is involved. For low-frequency senses, such as EARTH, ROW, and ROAD senses of *bank*, the approach does not seem to be very effective. That is not a problem specific to the adaptive approach, and all other approaches in the literature suffer from the same problem of data sparseness. Even with static knowledge acquired from a very large corpus, these senses were disambiguated at a considerably lower precision rate than other senses.

6 Related Work

There has been increasing interest in using a machine to identify the intended sense of a polysemous word in a given context. Recently, various approaches to WSD have been proposed in the natural language processing literature, and old ideas have been superseded by newer ones at a rapid rate. Central to these development efforts are the kind of contextual knowledge encoded and the way this knowledge is represented and acquired. In this section, we review the recent literature on WSD from the perspectives of different types of contextual knowledge and their representational schemes.

6.1 Lexicalized vs. Conceptual Encoding of Context

Any kind of scheme for acquiring contextual information of word sense must begin with a way of identifying the word sense since word sense is an abstract concept not clear on the surface. Once this is done, we can use the surrounding words to build a contextual representation of the word sense for WSD. There are three approaches to the chicken-and-egg problem of dividing word senses. First, one can resort to human intervention to get a hand-tagged corpus of word senses. Most early WSD works used this approach and went to the trouble of hand-tagging the intended sense of each polysemous word in the training corpus [Kelly and Stone 1975; Hearst 1991]. Second, one can take the numbered sense entries readily available in a machine-readable dictionary and treat their definitions and examples as contextual information [Lesk 1986; Veronis and Ide 1990; Wilks *et al.* 1990; Guthrie *et al.* 1991]. The third way of identifying word sense exploits linguistic constraints. For instance, three linguistic constraints can be exploited for successful sense tagging and WSD.

• **One sense per discourse** The senses of all instances of a polysemous word are highly consistent within any given document.

•**One sense per collocation** Nearby words provide strong and consistent clues to the sense of a target word, conditional on the relative distance, order, and syntactic relationship.

•**One sense per translation** Translations in a bilingual corpus can be used to represent the senses of words.

As an example of the first constraint, consider the word *suit*. The constraint captures the intuition that if the first occurrence of *suit* is a LAWSUIT sense, then later occurrences in the same discourse are also likely to refer to LAWSUIT [Gale, Church and Yarowsky 1992a]. The second constraint indicates that most works on statistical disambiguation have made the basic assumption that word sense is strongly correlated with certain contextual features, like occurrence of particular words in a window around the ambiguous word. However, Yarowsky [1995] proposed an approach in which strong collocations were identified for WSD. If a bilingual corpus was available, differences in translations of the polysemous word allowed one to delineate the intended sense, particularly in the case of contrasting polysemy. Gale,

Church and Yarowsky [1992b] used French translations in parallel texts to disambiguate some polysemous words in English. For instance, the senses of *duty* were usually translated as two different French words, *droit* and *devoir*, respectively, representing the senses *tax* and *obligation*. Thus, a number of *tax* sense instances of *duty* could be collected by extracting instances of *duty* that were translated as *droit,* and the same could be done for *obligation* sense instances of *duty*.

Once word senses are identified in one way or another, the context of a particular word sense can then be acquired and encoded in some way for use in the subsequent disambiguation step. There are at least two ways of encoding contextual knowledge. The obvious way, the lexicalized representation, is a surface scheme that keeps a weighted list of words appearing in the context of a particular sense. On the other hand, the conceptual representation encodes the classes of words that might appear in the context.

6.1.1 Lexicalized Representation of Context

Dictionary Definitions as Context Lesk [1986] described a word-sense disambiguation technique based on the number of overlaps between words in a dictionary definition and the fixed-size window of words surrounding the target. The author reported WSD performance ranging from 50% to 70% when the method was applied to a sample of ambiguous words. Lesk's method had failed to determine the correct senses of words when two or more senses of a word had the same number of overlaps with the context. Veronis and Ide [1990] constructed an artificial neural network from sense definitions, representing each word in the definition text as a node in the network. Different senses of each word competed with each other through the mechanism of spreading activation initiated at the nodes of contextual words. White [1988], Guthrie *et al.* [1991], and Slator [1991] used measures of words in context overlapping with dictionary definitions. One major problem of these earlier approaches was their lack of abstraction. The rich semantic information in the definition, such as the genus term, differentia, and implicit topics, was not exploited to the fullest.

Context as Co-occurrence Probabilities Gale, Church and Yarowsky [1992b] indicated that translation in a bilingual corpus could be used to provide tagged material for supervised learning of WSD knowledge. In their experiment, French translations were, in effect, used to represent the senses of some English words under the assumption of *one-sense-per-translation*. The Bayesian model was used to represent the contextual words in terms of their probabilities of occurrence. They reported a 90% accuracy rate in discriminating between two constrasting senses of six ambiguous nouns in the Canadian Hansards: *duty, drug, land, language, position, and sentence*. The weaknesses of this approach include the dreaded problem of data sparseness. Even when a very large corpus is available, it is still difficult to guarantee that each word sense will have enough contextual samples to avoid running into the problem of zero frequency, namely, the difficulty of assigning appropriate probabilistic values to words that do not appear in these contextual samples.

Smoothing Co-occurrence Probabilities Yarowsky [1992] improved on the WSD method proposed by Gale, Church and Yarowsky [1992b] by smoothing the concurrence probability via predefined semantic classification. Basically, that was done by lumping the probabilities related to all the senses in a thesaurus category to smooth the zero frequency cases. For instance, the contextual information of *bird* and other animals was used to build a contextual representation for all the senses in the animal category in Roget's Thesaurus [1987]. His experiment showed in a *close test* using Grolier's Encyclopedia that instances of twelve words, *bass, bow, cone, duty, galley, interest, issue, mole, sentence, slug, star,* and *taste*, could be disambiguated with an average precision rate of 92%. However, a very large corpus is required to train such a lexicalized contextual model, and clearly this kind of static model has a portability problem.

6.1.2 Conceptual Representation of Context

Context as Definition-Based Conceptual Co-occurrence Luk [1995] advocated using defining words in the MRD for the contextual representation of word sense. Reminiscent of an earlier work by Wilks *et al.* [1990], Luk proposed a definition-based concept co-occurrence model (DBCC) for WSD. With the model, the context of each word sense is represented using a vector of LDOCE defining words in the sense definition. The author argued that by using a fixed, relatively small number of concepts, a small corpus could provide enough concept co-occurrence data for statistical sense disambiguation. In a close test, the DBCC model trained on the Brown corpus was found to be capable of disambiguating 60%[9] of the instances of the same twelve ambiguous words used in Yarowsky's experiment.

Context as Thesaurus Categories Many researchers have exploited the semantic categories in a thesaurus, such as Roget's and LLOCE, or the subject information in a dictionary for context representation and WSD. Walker and Amsler [1986] applied subject codes in LDOCE as semantic representation for WSD. Black [1988] reported an accuracy rate of around 50% when Walker and Amsler's algorithm was applied to a sample of five ambiguous words: *interest, point, power, state,* and *term*. Pure conceptual representation is the most economical kind of WSD model since it requires the smallest parameter space and requires no substantial texts for training. Chen *et al.* [1996] proposed a mixed representational scheme for context based on contextual words as well as LLOCE topics. With a contextual representation acquired from example sentences in LDOCE/E-C, the authors reported that the method could disambiguate around 70% of the instances of thirteen polysemous words in the Brown corpus.

---

[9] The originally reported value, 77%, was based on the average of the precision rates for all twelve words. This form of evaluation is sensitive to the outcome of a handful of test samples since the precision rate of a word with a couple of samples could have an overly strong impact on the average. In this paper, we use the average rate of precision calculated over all instances.

6.2 Topical vs. Local Representation of Context

In almost all the studies described in Section 2.1, topical context was used in WSD. In a number of research works related to machine translation, researchers have used local context to solve a problem closely related to WSD, namely, the lexical choice problem. We will examine these two different kinds of contextual information in this section.

6.2.1 Topical Context

With topical representation of context, the context of a given sense of a target word is a bag of words without any structure. Information in topical context is generally quite helpful for WSD. For instance, consider Examples (16) and (17) extracted from the Brown corpus, each containing an instance of the ambiguous word *bass*.

(16) … for scintillating flights of meaningless improvisations, and he has a quiet way of getting back and restating the melody after the improvising is over. In this he is sticking with tradition, however far removed from it he may seem to be. SHEARING TAKES OVER George Shearing took over with his well disciplined group, a sextet consisting of vibes, guitar, **bass**, drums, Shearing's piano and a bongo drummer. He met with enthusiastic audience approval, especially when he swung from jazz to Latin American things like the Mambo. Shearing, himself, seemed to me to be playing better piano than in his recent Newport appearances. A very casual, pleasant program- one of those easy-going things that make Newport's afternoon programs such a …

(17) … Breakfast was at the Palace Hotel, luncheon was somewhere in the mountain forest, and dinner was either at Boulder Creek or at Santa Cruz. Gazing too long at the scenery could be tiring, so halts were contrived between meals. Then the Chinese hostler, who rode with Vernon on the box, would break open a hamper and produce filets of smoked **bass** or sturgeon, sandwiches, pickled eggs, and a rum sangaree to be heated over a spirit lamp. In spring and in autumn the run was made for a group of botanists which included an old friend of mine. They gathered roots, bulbs, odd ferns, leaves, and bits of resin from the rare Santa Lucia fir, which exists only on a forty-five mile strip on the westerly side of these mountains. In the Spanish …

Intuitively, the first instance of *bass* can be disambiguated as INSTRUMENT-bass since *guitar*, *drum*, *piano*, *jazz*, etc. are likely to appear in the topical context of INSTRUMENT-bass. Similarly, the second instance can be disambiguated as FISH-bass since *meal*, *sandwiches*, *egg*, etc. are often found in the topical context of FISH-bass. Generally, the sense representation of topical context is acquired from a very large corpus. Gale, Church and Yarowsky [1992b] experimented on acquiring topical context from a substantial bilingual training corpus and reported good results.

6.2.2 Local Context

Local context includes structured information about word order, distance, and syntactic features. For instance, the local context of *a line from* does not suggest the same sense for the word *line* as *a line for* does.

Trigram as Local Context Brown *et al.* [1990] used the trigram model as a way of resolving sense ambiguity for lexical selection in statistical machine translation. This model makes the assumption that only the previous two words have any effect on the translation, and thus, the word sense of the next word. The model was used to attack the problem of lexical ambiguity and produced satisfactory results, under some strong assumptions. For instance, the authors showed that the French sentence *Je vais prendre la decision* could be correctly translated as *I will make the decision* using this model. Although in isolation, *take* was more likely than *make* to translate as *prendre*, the trigram language reversed the decision in favor of *make*. A major problem with the trigram model is long distance dependency. For instance, the model incorrectly rendered the French sentence *Je vais prendre ma propre decision* as *I will take my own decision*. The language model did not consider *make my own decision* more probable since *prendre* and *decision* did not fall within a window of three words.

Lexical Relation Dagan, Itai and Schwall [1991] and Dagan and Itai [1994] made use of translations of different senses from a Hebrew/English bilingual dictionary to disambiguate contexts. Local context in the form of lexical relations was analyzed in a foreign corpus. The basic idea of the algorithm is best explained with an example. Given two Hebrew words *hoze* and *shalom*, *hoze* has two translations in English: *contract* and *treaty*, while *shalom* is often translated into English as *peace*. Their experiment showed that all instances of *peace* appear before *treaty* and none before *contract* in the corpus of English language. Therefore, the authors concluded that this instance of *hoze* in the phrase *hoze shalom* was best translated as *treaty*. The authors experimented on lexical choice with 105 Hebrew words and 54 German words from news articles. The precision rates achieved ranged from 75% to 92% for coverage rates between 59% and 70%.

Approximating Lexical Relation Brown *et al.* [1991] described a statistical algorithm for partitioning the senses of a word into two groups. The authors used mutual information to find a local contextual feature that most reliably indicated which of the senses of the French ambiguous word was used. For instance, for the verb *prendre*, the object was a good indicator: *prendre une measure* translated as *to take a measure,* and *prendre une decision* as *to make a decision.* Therefore, words (any word, first verb or first noun) immediately to the left or right of the word were evaluated for their effectiveness as good indicators for WSD and lexical choice. The authors reported 20% improvement in the performance of a machine translation system (from 37 to 45 sentences correct out of 100) when the words were first disambiguated in this way.

7 Conclusions

We have described an adaptive approach to word sense disambiguation. Under this new learning strategy, a contextual representation for each sense discriminator is first built based on the sense definition and example sentence in MRD and represented as a weighted-vector of concepts represented by word lists in a thesaurus. This knowledge representation acquired through MRD is based on a limited number of concepts; thus, the dreaded problem of data sparseness is avoided. Conceptual knowledge also offers the additional advantages of reduced storage requirements and increased efficiency due to reduced dimensionality. Also, we can correctly identify at least 50% of the word senses in unrestricted texts. In addition, these disambiguated texts can be used to adjust the fundamental knowledge in an adaptive fashion so to improve disambiguation precision. We have demonstrated that this approach can outperform established static approaches based on direct comparison of results obtained for the same words. This level of performance is achieved without lengthy training or the use of a very large training corpus.

Appendix A

A Glossary of LLOCE Topics

Here, we list 129 topics found in LLOCE. The column labeled "Topic" shows a set of two-character symbols representing the topics in LLOCE. Each topic is giving a gloss.

Topic	Glossary	Topic	Glossary
Aa	Life and living things	Bg	Bodily states and associated activities
Ab	Living creatures generally	Bh	Bodily conditions relating to health, sickness, and disability
Ac	Animals/Mammals	Bi	Diseases and ailments
Ad	Birds	Bj	Medicine and general medical care
Ae	Reptiles and amphibians	Ca	People
Af	Fish and other water creatures	Cb	Courting, sex, and marriage
Ag	Insects and similar creatures	Cc	Friendship and enmity
Ah	Parts of animals	Cd	Death and Burial
Ai	Kinds and parts of plants	Ce	Social organization in groups and places
Aj	Plants generally	Cf	Government
Ba	The body generally	Cg	Politics and elections
Bb	The body: overall	Ch	Political tension and trouble
Bc	The head and the face	Ci	Social classifications and situations
Bd	The trunk, arms, and legs	Cj	Law and order generally
Be	The skin, the complexion, and the hair	Ck	Courts of law and legal work
Bf	Fluids and waste products of the body	Cl	The police, security services, crime, and criminals

Topic	Glossary	Topic	Glossary
Cm	Prison and punishment	Gh	General grammatical words
Cn	Warfare, defence, and the army	Ha	Substances and materials generally
Co	The armed forces	Hb	Objects generally
Cp	Religion and beliefs	Hc	Specific substances and materials
Da	Architecture and kinds of houses and buildings	Hd	Equipment, machines, and instruments
Db	Parts of houses	He	Tools
Dc	Areas around and near houses	Hf	Containers
Dd	Residence	Hg	Electricity and electrical equipment
De	Belonging and owing, getting and giving	Hh	Weapons
Df	Furniture and household fittings	Ia	Making things
Dg	Clothes and personal belongings	Ib	Arts and crafts
Dh	Cleaning and personal care	Ic	Science and technology
Ea	Food generally	Id	Industry and work
Eb	Food	Ie	Education
Ec	Drinks	Ja	Numbers and quantities
Ed	Cigarettes and drugs	Jb	Mathematics
Ee	The preparation and quality of food	Jc	Measurement
Ef	Places and people associated with food and drink	Jd	Money
Eg	Farming	Je	Banking, wealth, and investment
Fa	Feeling and behavior generally	Jf	Commerce
Fb	Liking and not liking	Jg	Shopping and general expenses
Fc	Good and evil	Jh	Business, work, and employment
Fd	Happiness and sadness	Ka	Entertainment generally
Fe	Anger, violence, stress, calm, and quietness	Kb	Music and related activities
Fh	Kindness and unkindness	Kc	Recording sound, listening to the radio, etc.
Fi	Honesty, loyalty, trickery, and deceit	Kd	Drama, the theatre, and show business
Fj	Relaxation, excitement, interest, and surprise	Ke	Sport and games generally
Fk	Actions of the face related to feelings	Kf	Indoor games
Fl	Senses and sensations	Kg	Children‘s games and toys
Ga	Thinking, judging and remembering	Kh	Outdoor games
Gb	Knowing and learning	La	The universe
Gc	Communicating, mainly by speaking and talking	Lb	Light and color
Gd	Communicating, mainly by reading and writing, printing and publishing, radio and television	Lc	Weather and temperature
Ge	Communication and information	Ld	Geography
Gf	Language	Le	Time generally
Gg	Grammar	Lf	Beginning and ending

Topic	Glossary	Topic	Glossary
Lg	Old, new, and young	Nb	Possibility, chance, and necessity
Lh	Periods of time and their measurement	Nc	General, usual, unusual, etc.
Li	Grammatical words and phrases relating to time	Nd	Size, importance, and availability
Ma	Moving, coming, and going	Ne	Doing things
Mb	Putting and taking, pulling and pushing	Nf	Causing
Mc	Travel and visiting	Ng	Resemblance, difference, and change
Md	Vehicles and transport on land	Nh	Rightness, fairness, purpose, use, and strength
Me	Places	Ni	Fullness, heaviness, thickness, stiffness, roughness, etc.
Mf	Shipping	Nj	Actions and positions
Mg	Aircraft	Nk	Cutting, joining, breaking, and destroying
Mh	Location and direction	Nl	Showing, hiding, finding, saving, and similar words
Na	Being, becoming, and happening		

References

Benson, M., E. Benson and R. Ilson. *The BBI Combinatory Dictionary of English,* John Benjamins Publishing Company. Amstersam/Philadelphia. 1993.

Black, E. "An Experiment in Computational Discrimination of English Word Sense. " *IBM Journal of Research and Developmen.* Vol. 32. pp. 185-194. 1988.

Brill, E. *A Corpus-Based Approach to Language Learning,* PH. D. thesis. Department of Computer and Information Science. University of Pennsylvania. 1993.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer and P. S. Roosin. "A Statistical Approach to Machine Translation." *Computational Linguistics*,.16(2). pp. 79-85. 1990.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra and R. L. Mercer. "Word–Sense Disambiguation using Statistical Methods" In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistic.* pp. 264-270. 1991.

Chang, J. S., R. H. Shu and M. H. Chen. "Automatic Extraction Rules on Preposition Phrase." In *Proceedings of R.O.C. Computational Linguistics Conference IX* (*ROCLING-IX*). pp. 295-320. 1996.

Chen, J. N. and J. S. Chang, "A Concept-based Adaptive Approach to Word Sense Disambiguation. " In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics*. pp. 237-244. Montreal. Canada. 1998a.

Chen, J. N. and J. S. Chang. "TopSense: A Topical Sense Clustering Method based on Information Retrieval Techniques on Machine Readable Resources." *Special Issue on Word Sense Disambiguation*. *Computational Linguistics*. 24(1). pp. 61-95. 1998b.

Chen, J. N., J. S. Chang, H. H. Sheng and S. J. Ker. "Combining Machine Readable lexical Resources and Bilingual Corpora for Broad Word Sense Disambiguation." In *Proceedings of the Second*

Conference of the Association for Machine Translation. pp. 115-124. Montreal. Quebec. Canada. 1996.

Dagan, I. and A. Itai. "Word Sense Disambiguation Using a Second Language Monolingual Corpus." *Computational Linguistic.* 20(4). pp. 563-596. 1994.

Dagan, I., A. Itai, and U. Schwall. "Two Languages are More Informative than One." In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. pp. 130-137. 1991.

Gale, W. A., K. W. Church and D. Yarowsky. "One Sense Per Discourse," In *Proceedings of the Speech and Natural Language Workshop*. pp. 233-237. 1992a.

Gale, W. A., K. W. Church, and D. Yarowsky. "Using Bilingual Materials to Develop Word Sense Disambiguation Methods." In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*. pp. 101-112. 1992b.

Guthrie, J., L. Guthrie, Y. Wilks and H. Aidinejad. "Subject-dependent Co-occurrence and Word Sense Disambiguation." In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. pp. 146-152. 1991.

Hearst, M. "Noun Homonym Disambiguation using Local Context in Large Text Corpora." In *Proceedings of the 7th International Conference on of UW Centre for the New OED and Text Research: Using Corpora*. pp. 1-22. 1991.

Hawking, D. and P. Thistlewaite. "Proximity Operators – So Near and So Far," In *Proceedings of the fourth Text REtrieval Conference* (*TREC-4*). pp. 1-13. 1995.

Kelly, E. and P. Stone. *Computer Recognition of English Word Senses*. North-Holland. Amsterdam. 1975.

Leacock, C., G. Towell and E. M. Voorhees. "Corpus-based Statistical Sense Resolution." In *Proceedings of the ARPA Workshop on Human Language Technology.* 1993.

Lesk, M. E. "Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone," In *Proceedings of the ACM SIGDOC Conference.* pp. 24-26, Toronto. Ontario. 1986.

Luk, A. K. "Statistical Sense Disambiguation with Relatively Small Corpora using Dictionary Definitions." In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. pp. 181-188. 1995.

McArthur, T. *Longman Lexicon of Contemporary English.* Longman Group (Far East) Ltd.. Hong Kong, 1992.

Proctor, P. (ed.) *Longman Dictionary of Contemporary English*. Harlow: Longman Group. 1978.

Proctor, P. (ed.) *Longman English-Chinese Dictionary of Contemporary English*. Longman Group (Far East) Ltd.. Hong Kong. 1988.

Roget's Thesaurus of English words and Phrases. Longman Group UK Limited. 1987.

Slator, B. "Using Context for Sense Preference." In Zernik (ed.) *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon.* Lawrence Erlbaum. Hillsdale. NJ. 1991.

Veronis, J. and N. Ide. "Word Sense Disambiguation with Very Large Neural Networks Extracted from Machine Readable Dictionaries." In *Proceedings of the 13th International Conference on Computational Linguistics*. pp. 389-394. 1990.

Walker, D. and R. Amsler. "The use of Machine-Readable Dictionaries in Sublanguage Analysis." In *Analyzing Language in Restricted Domains*. Grishman. R. and R. Kittredge (eds.). Lawrence Erlbaum Associates. Hillsdale. New Jersey. 1986. (also available in R. Kittredge (ed.), *Proceedings of Workshop on Sublanguage Analysis*; New York 1984).

White, J. S. "Determination of Lexical-Semantic Relations for Multi-Lingual Terminology Structures." In *Relational Models of the Lexicon*. Cambridge University Press, Cambridge. UK. 1988.

Wilks, Y. A., D. C. Fass, C. M. Guo, J. E. McDonald, T. Plate and B. M. Slator. "Providing Tractable Dictionaty Tools." *Machine Translation*. 5. pp. 99-154. 1990.

Yarowsky, D. "Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora." In *Proceedings of the 14th International Conference on Computational Linguistics*. pp. 454-460. Nantes. France. 1992.

Yarowsky, D. "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods." In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*. pp. 189-196. 1995.

Zipf, G. *Human Behavior and the Principle of Least Eeffect*. Hafner. New York. 1994.

Computational Linguistics and Chinese Language Processing
Vol. 5, No. 2, August 2000, pp. 43-58

Design and Evaluation of Approaches to Automatic Chinese Text Categorization

Jyh-Jong Tsay and Jing-Doo Wang*

Abstract

In this paper, we propose and evaluate approaches to categorizing Chinese texts, which consist of term extraction, term selection, term clustering and text classification. We propose a scalable approach which uses frequency counts to identify left and right boundaries of possibly significant terms. We used the combination of term selection and term clustering to reduce the dimension of the vector space to a practical level. While the huge number of possible Chinese terms makes most of the machine learning algorithms impractical, results obtained in an experiment on a CAN news collection show that the dimension could be dramatically reduced to 1200 while approximately the same level of classification accuracy was maintained using our approach. We also studied and compared the performance of three well known classifiers, the Rocchio linear classifier, naive Bayes probabilistic classifier and k-nearest neighbors(kNN) classifier, when they were applied to categorize Chinese texts. Overall, kNN achieved the best accuracy, about 78.3%, but required large amounts of computation time and memory when used to classify new texts. Rocchio was very time and memory efficient, and achieved a high level of accuracy, about 75.4%. In practical implementation, Rocchio may be a good choice.

Keywords: Term Clustering, Term Selection, Text Categorization.

1. Introduction

In recent years, we have seen a tremendous growth in the number of online text documents available on the Internet, in digital libraries and news sources. Effective location of information in these huge resources is difficult without good indexing as well as organization of text collections. Automatic text categorization, which is defined as the task of assigning predefined class (category) labels to free text documents, is one of

* Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan 62107, ROC. E-mail: {tsay,jdwang}@cs.ccu.edu.tw

the main techniques that are useful both in organizing and in locating information in these huge collections.

Many approaches to text categorization and web page classification have been proposed [2,9,12,20]. Most of them have been evaluated using English texts. Evaluation of these approaches using texts in Chinese and other oriental languages has been limited. In [22], Yang *et al.* proposed and evaluated several approaches to Chinese text categorization. The number of training and testing documents used was 2306, and the number of extracted terms used was 4711. Yang's work was quite preliminary and achieved classification accuracy of only about 67%. Since then, tremendous advances have been made in categorization techniques [20]. Most of the recently proposed techniques have not been evaluated using Chinese texts.

The objective of this study was to design and evaluate approaches to categorizing Chinese texts. In particular, we implemented and evaluated approaches which consist of the following processes: term extraction, term selection, term clustering and text classification. Note that in Chinese texts, although a sentence is composed of a sequence of terms, no white spaces are inserted to separate terms from each other. Term extraction which segments sentences into term sequences is a difficult task [5]. Several approaches have been proposed to extract terms from Chinese texts [4,13]. In this paper, we propose a scalable approach [17] which is based on String B-trees proposed in [7] and is capable of handling huge numbers of text documents. Our approach uses frequency counts to identify possible term boundaries as proposed in [13] and is able to identify new terms which occur very often in Chinese texts.

However, the number of terms in Chinese can be very large. It is very easy to encounter 10^6 or even more terms in moderately-sized collections. The huge number of possible terms results in very high dimensionality when documents was presented in a vector space model and makes many machine learning algorithms impractical. To reduce the dimension to a practical level, we propose to perform term selection and term clustering on extracted terms. In particular, we use the χ^2 statistic[16] to select terms that are highly correlated to class categories. In [16], we presented an extensive comparison of several measures for term selection in Chinese text categorization, such as the odds ratio, information gain, mutual information, and χ^2 statistic. Experimental results shows that the χ^2 statistic approach achieves the best performance. Notice that in term selection, if only a small number of terms is selected, a document may contain very few or even none of the selected terms, and thus will be classified into the default class. On the other hand, a large number of selected terms make automatic categorization computationally impractical. We thus allow a large number of terms to be selected and then perform term clustering to group similar terms into clusters.

A large number of algorithms for clustering are Available [11]. Most of them are unsupervised and ignore any class labels that are given. In this study, we used distributional clustering [2], which explicitly takes advantage of the class labels to group terms with similar class distributions into the same cluster. In an experiment on a collection of CNA news [1] articles, the number of terms extracted was 548363.

Experimental results show that the level of classification accuracy could be maintained while the dimension was reduced to 1200 by selecting 90000 terms first and then clustering them into 1200 clusters. Notice that term selection and term clustering also can compeensate for imprecision in term extraction as erroneous terms can be dropped out during term selection or grouped with more significant terms through term clustering. In addition to term selection and term clustering algorithms, there are others which can be applied to reduce the level of dimensionality, such as Principle Component Analysis (PCA) [6]. PCA is an unsupervised dimensional reduction technique, whereas distributional clustering is supervised and can take advantage of class labels to concentrate effort on the specific task of categorization. We expect distributional clustering to perform well in the context of text categorization.

In this paper, we also compare extensively three well-known classifiers, including the Rocchio linear classifier [12], naive Bayes probabilistic classifier, and k-nearest neighbor (kNN) classifier [20]. We observed in an experiment that the classification accuracy of Rocchio and kNN improved slightly as the dimension was reduced to 1200 by means of term selection and term clustering but that the accuracy of the naive Bayes classifier dropped slightly. This might have been due to the fact that term clustering refines the shapes of each cluster but distorts the distribution of each term. Overall, kNN achieved the best accuracy, about 78.3%, but required large amounts of computation time and memory when used to classify new texts. Rocchio is very time and memory efficient, and achieves accuracy of about 75.4%, which is slightly worse than kNN.

Recently, Huang *et al.* [10] evaluated the weight matrix approach, which estimates the relative importance of the keywords in each class and classifies a test news to the class that maximizes the sum of the weights of keywords appearing in that news. Although they achieved about 88% classification accuracy, their experiment was different from ours as well as those used in much related research [3,22]. First, the training news did not come from the same news source as the test news, but come from a thesaurus [19] that was carefully built by linguistic specialists. Second, the test news was classified by readers who could employ logic that was close to that assumed by the classification algorithms but different from that employed by the editors. Third, a piece of test news could be assigned to multiple classes when it covered topics from different classes. In fact, for a collection of 1136 news items, 1380 class labels were assigned, which indicates that about 20% of the test news iteems had multiple class labels. However, in the CNA news collection used in this study, each news items had exactly one predefined class no matter how many topics it covered. It is not clear whether or not the weight matrix approach can achieve the same performance when all the differences are removed.

The remainder of this paper is organized as follows. Section 2 sketches the String B-tree approach to term extraction. Section 3 describes the χ^2 statistic approach to term selection. Section 4 describes distributional clustering. Section 5 reviews the classifiers compared in this paper. Section 6 gives experimental results. Section 7 gives conclusions.

2. Term Extraction

In this paper, we propose a scalable approach [18] to term extraction, which is based on String B-trees (SB-trees) [7]. This approach can handle large text collections and can identify newly created terms frequently found in Chinese. It does not use a dictionary but rather uses frequency counts to identify the boundaries of possible terms as in [13]. We will describe the term extraction method in the following.

Let *w* be a string. For any character *x*, let *P (wx|w)* be the probability that w is followed by *x*, and let *P(xw|w)* be the probability that w is preceded by *x*. We say that w passes right boundary verification if *P(wx|w)* < θ_1 for all x and passes left boundary verification if *P(xw|w)* < θ_2 for all *x*. The probability *P(wx|w)*, resp. *P(xw|w)*, is estimated by $\frac{TF(wx)}{TF(w)}$, resp. $\frac{TF(xw)}{TF(w)}$, where *TF(y)* is the term frequency of string y. String *w* is identified as a significant term if it passes both right and left boundary verifications. In this paper, we simply set $\theta_1 = \theta_2 = 1$, which means that w will be identified as a significant term when it has at least two distinct successor and predecessor characters. For each class, we build two SB-trees, one for all the suffixes [8] of the original texts used for right boundary verification, and the other for the suffixes of the reversed texts which is used for left boundary verification. Notice that SB-trees are scalable; they can maintain dynamic collections and identify new terms as new articles are inserted.

3. Term Selection

Term selection is performed to choose representative terms for each class such that these terms can distinguish one class from the others. After the term extraction process is completed, there are many terms remain that are not informative for categorization. In [16], we extensively compared several measures used for term selection in Chinese text categorization, such as the odds ratio, information gain, mutual information, and χ^2 statistic. Experimental results show that the χ^2 statistic approach achieves the best performance when combined with the naive Bayes classifier. In this study, we used the χ^2 statistic [21] approach to perform term selection.

For a term *t* and a class *c*, the χ^2 statistic measures the correlation between *t* and *c*. Let *A* be the number of times *t* and *c* co-occur, let *B* be the number of times *t* occurs without *c*, let *P* be the number of times *c* occurs without *t*, let *Q* be the number of times neither *t* nor *c* occur, and let *N* be the total number of documents. The χ^2 statistic is defined as $\chi^2(t,c) = \frac{N \times (AQ - BP)^2}{(A+P) \times (B+Q) \times (A+B) \times (P+Q)}$.

Notice that the χ^2 statistic approach prefers terms that are highly correlated with a particular class. For each term, the χ^2 statistic scores with regard to different classes can be different. In [21], Yang used the

average or the maximum of the scores to select representative terms, which may result in a biased distribution of selected terms between classes. To avoid this situation, we select from each class the same number of terms having the largest χ^2 statistic in that class.

4 Term Clustering

We perform term clustering to further reduce the dimension of the vector space after the term selection process. In order to avoid the situation in which a document contains none of the selected terms, in term selection, we select a suitable large set of terms which may require a large amount of computation time and memory for classification. Term clustering groups similar terms into one cluster that no longer distinguishes between constituent terms. In this study, we used distributional clustering [2], which groups terms with similar distributions over classes into the same cluster. Note that distributional clustering can compensate for the drawback of term extraction, where incomplete terms are clustered into the group containing their original terms. On the other hand, when training data is sparse, performance may be improved by averaging statistics of similar words together so that the resulting estimates are more robust. We describe distributional clustering [2] in more detail in the following.

Term clustering algorithms define a similarity measure between terms and group similar terms into term clusters. In distributional clustering, the difference between two term distributions is measured by Kullback-Leibler (KL) divergence. For term t_i and term t_j, the KL divergence, denoted as $D(P(C|t_i)\|P(C|t_j))$, is defined as $-\sum_{k=1}^{|C|} P(C_k|t_i)\log\frac{P(C_k|t_i)}{P(C_k|t_j)}$, where $|C|$ is the number of classes and $P(C_k|t_i)$ is the probability of class C_k given term t_i. To avoid the odd properties of KL divergence, such as asymmetry, we use the average KL divergence defined as $\frac{P(t_i)}{P(t_i \vee t_j)} \cdot D(P(C|t_i)\|P(C|t_i \vee t_j)) + \frac{P(t_j)}{P(t_i \vee t_j)} \cdot D(P(C|t_j)\|P(C|t_i \vee t_j))$, where $t_i \vee t_j$ represents clustering of term t_i and term t_j into one group. Based on the average KL divergence, we apply a simple greedy agglomerative algorithm to cluster terms as follows. Let *M* be the number of final clusters. Initially, *M* terms are selected as seeds. Each term represents a singleton cluster. The following process is repeated until all the terms have been added: the two most similar clusters are merged into one cluster, and then the term that has the highest χ^2 statistic measure among the remaining terms is added as a singleton cluster. The initial *M* seeding terms are uniformly selected from all classes. That is, from each class, the $M/|C|$ terms that have the highest χ^2 statistic measure are selected as initial seeds. This avoids the problem of bias [2], where the *M* initial clusters may prefer some classes.

5. Classifiers

In this paper, we compare three wellknown classifiers, including the Rocchio linear classifier, naive Bayes (NB) probabilistic classifier and k-nearest neighbor (kNN) classifier, which are reviewed in the following sections.

5.1 Rocchio Linear Classifier

The Rocchio algorithm is a training algorithm [12] for linear classifiers and was initially developed for information retrieval in the vector space model. The basic idea is to construct one prototype vector per class, using a training set of documents. Given a class, the training document collection consists of positive and negative examples. Positive examples are those documents belonging to that class, while negative examples are those documents not belonging to that class. The prototype vector of a class is the centroid of positive examples, tuned using negative examples. Let D_i be a document in the training collection *D*, represented as a vector $(d_{i,1}, d_{i,2}, \cdots, d_{i,n})$, where $d_{i,j}$ is the weight assigned to the *j*th term and *n* is the dimension of the document space. To determine $d_{i,j}$, we use the TF-IDF weighting method [15], which has been shown to be effective when used in the vector space model. Let $tf_{i,j}$ be the term frequency of the *j*th term in document D_i, and let df_j be the document frequency of the *j*th term in training collection *D*. In this paper, the TF-IDF weight is defined as $d_{i,j} = \log_2(tf_{i,j} + 1) * \log_2(\frac{N}{df_j})$, where *N* is the total number of documents in the training collection.

The prototype vector $G_i = (g_{i,1}, g_{i,2}, \cdots, g_{i,n})$ of class C_i is defined as $G_i = \frac{\sum_{D_i \in C_k} D_i}{|C_k|} - \eta \frac{\sum_{D_i \in D - C_k} D_i}{|D - C_k|}$, where η is the parameter that adjusts the relative impact of positive and negative examples. We have experimented with different values for η, including 0.25, 0.5, 0.75 and 1. The best choice of η in our experiment was found to be 0.5 when $n = 90000$, and was 0.25 when $n = 1200$. To classify a request document *X*, we compute the cosine similarity between *X* and each prototype vector G_i, and assign to *X* the class whose prototype vector has the highest degree of cosine similarity with *X*. Cosine similarity is defined as $CosSim(X, G_i) = \frac{\sum_{j=1}^{n} x_j \cdot g_{i,j}}{\sqrt{\sum_{j=1}^{n} x_j^2} \sqrt{\sum_{j=1}^{n} g_{i,j}^2}}$.

5.2 Naive Bayes (NB) Classifier

The Naive Bayes (NB) probabilistic classifiers have been studied for application to machine learning [14]. The basic idea in NB is to use the joint probabilities of terms and classes to estimate the probabilities of classes given a document. The naive part is the assumption of term independence, i.e., the conditional probability of a term, given a class, is assumed to be independent from the conditional probabilities of other words given that class. This assumption makes computation for NB classifiers far more efficient than that for the non-naive Bayes approaches [20] whose time complexity are exponential.

Let X be a request document; NB assigns to X the most probable class C_{NB} defined as $C_{NB} = \arg\max_{c_k \in c} P(C_k \mid X)$. By Bayes' theorem, $P(C_k \mid X) = \frac{P(X \mid C_k)P(C_k)}{\sum_{c_i \in C} P(X \mid C_i)P(C_i)}$. Due to the assumption of term independence, $P(X \mid C_k) = \Pi_{j=1}^{|X|} P(t_j \mid C_k)$, where $P(t_j \mid C_k)$ is the conditional probability of term t_j given class C_k. Notice that the above equation works well when every term appears in every document. However, the product becomes 0 when some terms do not appear in the given document. We use $P(t_j \mid C_k) = \frac{1 + TF(t_j, C_k)}{|T| + \sum_j^{|T|} TF(t_j, C_k)}$ in order to approximate $P(t_j \mid C_k)$ to avoid the possibility that the product will become 0, where $TF(t_j, C_k)$ is the frequency of occurrence of term t_j in documents of class C_k and $|T|$ is the total number of distinct terms used in the domain of document representation. The formula used to predict the probability of class value C_k for a given document X is $P(C_k \mid X) = \frac{P(C_k)\Pi_{t_j \in X} P(t_j \mid C_k)^{TF(t_j, X)}}{\sum_i P(C_i)\Pi_{t_j \in X} P(t_j \mid C_i)^{TF(t_j, X)}}$.

5.3 k-Nearest Neighbor (kNN) Classifier

Given an arbitrary request document *X*, kNN ranks its nearest neighbors among the training documents and uses the classes of the *k* top-ranking neighbors to predict the classes of *X*. The similarity score of each neighbor document when it is compared to *X* is used as the weight of the class of the neighboring document, and the sum of the class weights over the *k* nearest neighbors is used to perform class ranking[20].

In a kNN algorithm, each training document D_i as well as the request document X are represented by means of vectors as $(d_{i,1}, d_{i,2}, \cdots, d_{i,n})$ and $(x_1, x_2, \cdots, x_n)$, respectively. To

conduct categorization, the cosine similarity between each D_i and X is calculated. The training documents are sorted using the cosine similarity metric in descending order. Then the k top-ranking documents are selected. The final score of the request document X when compared to each class is calculated by summing the cosine similarity metric of these k selected documents and their class association. The class with the highest score is assigned to X. We have performed an experiment using different values of k, including 5, 10, 15, 20, 30, 50, 100, 150, 200 and 300. The best choice of k in our experiment is 15 when n = 90000 and is 10 when n = 1200.

6. Experimental Results

In our experiment, we used Chinese news articles from the Central News Agency (CNA)[1]. We used news articles spanning a period of one year, from 1/1/1991 to 12/31/1991, to extract terms. News articles from the six-month period 8/1/1991 to 1/31/1992 were used as training data to train classifiers. The testing data consisted of news articles from the one-month period 2/1/1992 to 2/28/1992. All the news articles were preclassified into 12 classes, listted in Figure 1. Note that the number of texts used was far larger than that employed in previous related researches [10,22]. As a result, the conclusions drawn based on our experimental results are believed to be more reliable.

		Train	Test
	CNA News Group	1991/8-1992/1	1992/2/1-2/28
政治	cna.politics.*	13482	1225
經濟	cna.economics.*	5768	776
文教	cna.edu.*	3203	379
社會	cna.judiciary.*	3132	492
體育	cna.l*	2778	415
財政	cna.finance.*	1958	151
軍事	cna.military.*	1818	261
交通	cna.transport.*	1801	279
股市	cna.stock.*	1779	200
社福	cna.health-n-welfare.*	1738	305
農業	cna.argriculture.*	1496	238
宗教	cna.religion.*	707	74
	Total	**39660**	**4795**

Figure 1 *The distribution of CAN news articles.*

The news articles were not uniformly distributed over the classes, as shown in Figure 1. We, thus, measure the classification accuracy at both micro and macro levels. Three performance measures were used to evaluate the performance of each classifier: *MicroAccuracy*, *MacroAccuracy* and

AccuracyVariance. Let |*C*| be the number of predefined classes, and let $|C_i|$ be the number of testing news articles that are preclassified into the *i*th class, and let $N = \sum_{i=1}^{i=|C|} |C_i|$ be the total number of testing news articles. Let $|H_{i,j}|$ be the number of testing news articles in C_i that are classified into C_j. Let $Acc(i) = \frac{|H_{i,i}|}{|C_i|}$ be the classification accuracy within class C_i. MicroAccuracy is defined as $\frac{\sum_{i=1}^{i=|C|} |H_{i,i}|}{N}$, which represents the overall average of classification accuracy. MacroAccuracy is defined as $\frac{\sum_{i=1}^{i=|C|} Acc(i)}{|C|}$, which represents the average of the classification accuracy within classes. AccuracyVariance is defined as $\frac{\sum_{i=1}^{i=|C|} (Acc(i) - MacroAccuracy)^2}{|C|}$, which represents the variance of accuracy among classes.

	n=90000	n=1200
Rocchio	00:21:13	00:04:26
naive Bayes	00:14:17	00:08:01
kNN	02:39:25	01:54:22

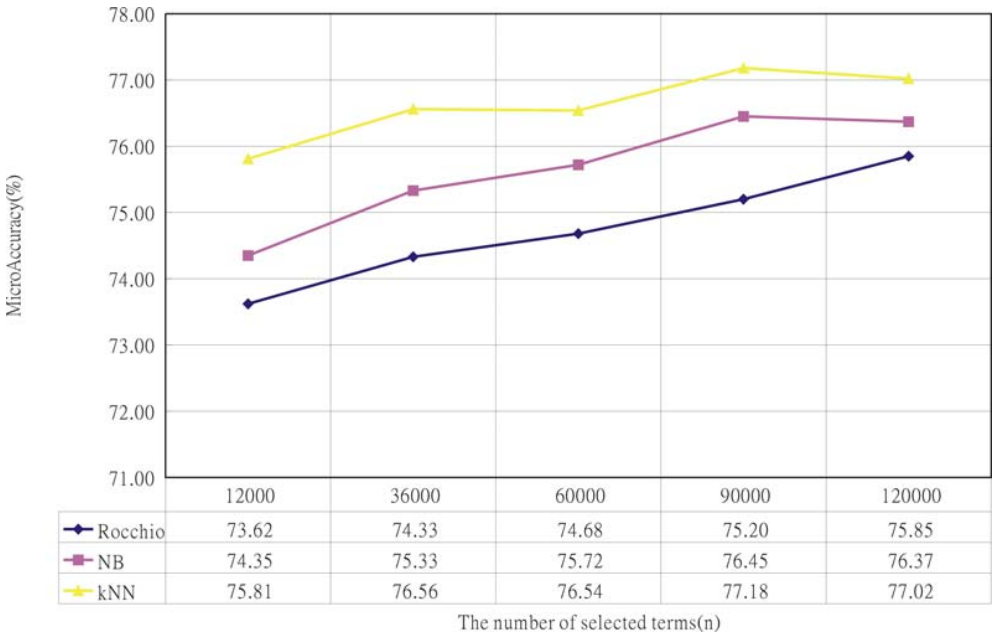
Figure 2 *Classification time.*

6.1 Dimension Reduction

We performed term extraction, term selection and term clustering to reduce the dimension. Both the space and time required to classify new documents could be reduced as the dimension of the vector space was reduced. Figure 2 shows the time needed to classify new documents, measured on a PC with a Pentium II 233 CPU, 128MB RAM and an IDE HardDisk, for dimension n = 90000 and 1200, respectively.

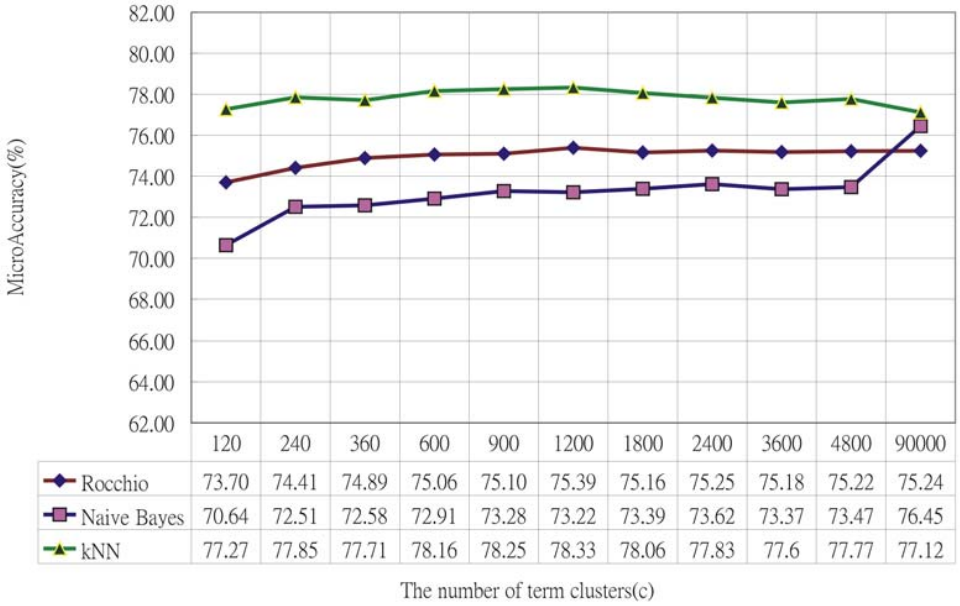
In the term extraction process, terms that appeared fewer than 10 times or in only one document were dropped out. We then used frequency counts to identify significant terms. The number of significant terms extracted was 548363. Term selection was then performed to select a subset of most representative terms. In order to find an appropriate number *p* of selected terms, we experimented for different values of *p*, including 12000, 36000, 60000, 90000 and 120000. We choose a *p* value of 90000 because kNN and NB achieved the best MicroAccuracy results of 77.12% and 76.45%, respectively, when *p* was 90000, as indicated in Figure 3. The selected terms were clustered using distributional clustering into term clusters. To choose a suitable number *c* of term clusters, we experimented with different values of *c*, including 120,

240, 360, 600, 900, 1200, 1800, 2400, 3600 and 4800. We choose a *c* value of 1200 because kNN and Rocchio achieved the best performance when *c* was 1200, as shown in Figure 4.



	12000	36000	60000	90000	120000
Rocchio	73.62	74.33	74.68	75.20	75.85
NB	74.35	75.33	75.72	76.45	76.37
kNN	75.81	76.56	76.54	77.18	77.02

Figure 3 *MicroAccuacy comparison(term selection).*



	120	240	360	600	900	1200	1800	2400	3600	4800	90000
Rocchio	73.70	74.41	74.89	75.06	75.10	75.39	75.16	75.25	75.18	75.22	75.24
Naive Bayes	70.64	72.51	72.58	72.91	73.28	73.22	73.39	73.62	73.37	73.47	76.45
kNN	77.27	77.85	77.71	78.16	78.25	78.33	78.06	77.83	77.6	77.77	77.12

Figure 4 *MicroAccuacy comparison(term clustering).*

Cluster ID(CID)

	12	100	207	225	300
1	二屆國	公路和	交響	犯案	今天在東京
2	二屆國代	在交通	交響樂團	刑事警察	交易所
3	二屆國代選舉	的快樂	巡迴演出	在逃	券交易所
4	的候選人	的班	的音樂	收押	證券交易所
5	候選人	旅行業	的舞	判處死刑	
6	候選人的	旅客的	奏會	官認為	
7	國大代表	旅遊協會	國立藝	押回	
8	國代候選人	泰航	國樂	前科	
9	國代選舉	機票	演奏	看守所	
10			演奏會	書指出	
11			舞蹈	處死刑	
12			樂家	被告	
13			樂團	槍枝	
14			鋼琴	辦案	
15			藝術學	警方在	

Figure 5 *Term clustering examples.*

		政治	經濟	交通	文教	體育	社會	股市	軍事	農業	宗教	財政	社福
CID=12	二屆國	3125	46	22	84	24	152			10	238	25	20
	二屆國代	1737	25	5	31	7	89			6	96	6	6
CID=100	旅行業	22		86		26	14						
	旅遊協會	31		74		15							
CID=300	券交易所	25	37					293				92	
	證券交易所	21	32					178				91	

Figure 6 *Term frequencies in each class.*

Figure 5 shows some examples of term groups. In addition to clustering similar terms to reduce the dimension, term clustering can also cluster redundant substrings that are erroneously identified during term extraction into the group that contains their original terms. For example, as shown in Figure6,"二屆國" and "二屆國代" are clustered into group 12; "證券交易所" and "券交易所" are clustered into group 300. On the other hand, the averaging statistics of similar words may result in more robust estimates. For example, "旅行業"(a travel agent) and "旅遊協會"(a travel agency association) are similar words and are clustered into group 100.

In [2], Baker claimed that performance can be improved by means of term clustering when training data is sparse because by averaging statistics of similar words, more robust estimates can be obtained. This was confirmed by our experiment. Note that our training data was quite sparse as the average number of none-zero items in training vectors was 106 when n is 90000, and was 79 when c is 1200. The memory space could be reduced by 25%($\frac{106-79}{106}=0.25$), and the averaged statistics of terms were more robust estimates when the percentage of none-zero items increased from 0.12%(=106/90000) to 6.58%(=79/1200) due to term clustering.

6.2 Classifiers Comparison

Overall, kNN achieved the best MicroAccuracy results, and Rocchio achieved slightly worse results, as shown in Figure 7 and Figure 8. Note that the MicroAccuracy results for Rocchio and kNN improved slightly from 75.24% and 77.12% to 75.39% and 78.33%, respectively, when the dimension of the vector space was reduced from 90000 to 1200 by means of distributional clustering. However, the performance of naïve Bayes dropped when terms were clustered. This might have been due to the fact that naive Bayes is more sensitive to term distributions which might be distorted by term clustering.

	Rocchio		NB		kNN	
	Recall	Precision	Recall	Precision	Recall	Precision
政治	72	81	77	80	88	72
經濟	71	78	76	75	73	79
文教	78	67	78	67	72	72
社會	73	90	75	87	72	92
體育	71	91	74	89	81	84
財政	83	60	83	60	89	54
軍事	82	53	77	56	45	76
交通	80	70	75	72	76	81
股市	96	91	93	95	96	96
社福	78	80	75	82	72	85
農業	76	76	72	84	78	80
宗教	62	32	47	43	45	60

Figure 7 *Recall(%)/precision(%) comparison(n=90000).*

	Rocchio		NB		kNN	
	Recall	Precision	Recall	Precision	Recall	Precision
政治	70	83	71	78	83	79
經濟	71	78	76	72	71	80
文教	77	68	78	61	74	71
社會	76	88	72	85	76	91
體育	75	91	67	92	82	85
財政	87	52	76	65	90	53
軍事	84	56	77	54	62	70
交通	79	71	73	70	83	79
股市	96	92	88	97	97	93
社福	81	79	73	81	79	83
農業	75	75	68	80	77	74
宗教	64	32	57	28	50	56

Figure 8 *Recall(%)/precision(%) comparison(n=1200).*

kNN prefered large classes as its MacroAccuracy result, 73.88%, was the lowest, but its MicroAccuracy result, 77.12%, was the best, as indicated in Figure 7. For highly related classes, kNN may prefer a larger class as the probability that the *k* nearest neighbors will belong to the larger class is higher. kNN achieved much better recall results than Rocchio for the class Politics (政治), which was the largest class in our news collections. However, Rocchio achieved much better recall results than kNN did for the

class Military (軍事). Note that the class Politics (政治) and the class Military (軍事) were highly correlated, as observed in [17], and that the class Politics (政治) was 5 times larger than the class Military (軍事).

In practical implementation, Rocchio could be a good choice. Rocchio is quite time and memory efficient because the time and memory requirements for the classification process are proportional to the number of classes. However, the time and memory requirements for kNN are proportional to the number of training documents. Rocchio is more noise tolerant than kNN and NB, as shown by the fact that the performance of kNN and NB worsened but the performance of Rocchio improved when n was changed from 90000 to 120000, as shown in Figure 3. Rocchio produced slightly worse MicroAccuracy results than kNN did, but can be improved to produce results approaching the performance of kNN by taking more than one representative to represent each class in [17].

7. Conclusions

In this paper, we have proposed and evaluated approaches to categorizing Chinese texts, which consist of term extraction, term selection, term clustering and text classification. For term extraction, we have proposed an approach based on String B-trees. It is scalable and is capable of handling very large numbers of text collections. We use the χ^2 statistic to perform term selection and use distributional clustering to perform term clustering to reduce the dimension of the vector space. Although many redundant terms are identified as significant terms during the term extraction process, the combination of term selection and term clustering somehow can compensate for this drawback by either filtering them out or clustering them into the group containing their original terms. Results of an experiment on a CNA news collection shows that the dimension could be reduced from 90000 to 1200 while approximately the same level of classification accuracy was maintained. We have also studies and compared the performance of three well known classifiers, the Rocchio linear classifier (Rocchio), naive Bayes (NB) probabilistic classifier and k-nearest neighbors (kNN) classifier, when they were applied to categorize Chinese texts. Overall, kNN achieved the best accuracy, about 78.3%, but required large amounts of computation time and memory to classify new texts. Rocchio was very time and memory efficient, and achieved accuracy of about 75.4%. In practical implementation, Rocchio may be a good choice. In addition, we have recently shown [17] that the performance of the Rocchio linear classifier can be improved to approximate that of kNN by taking multiple representative vectors to represent one class.

Acknowledgements

We would like to thank Dr. Chien, Lee-Feng and Mr. Lee, Min-Jer for kind help in gathering the CNA news articles.

References

[1] Central News Agency. http://www.can.com.tw/index.html

[2] Douglas Baker and Kachites McCallum. "Distributional clustering of words for text classification." *In Proceedings of the 21th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR'98), pages 96-103. 1998.

[3] Chen Chun-Liang and Lee-Feng Chien. "PAT-tree-based online corpus collection and classification." *In The Fourth International Workshop on Information Retrieval with Asian Languages*(IRAL'99), pages 78-82. 1999.

[4] Chien Lee-Feng. "PAT-Tree-Based keyword extraction for Chinese information retrieval." *In Proceedings of the 20th Ann Int ACM SIFIR Conference on Research and Development in Information Retrieval*(SIGIR'97), pages 50-58. 1997.

[5] Chien Lee-Feng and Hsiao-Tieh Pu. "Important issues on Chinese information retrieval." *In Computation Linguistics and Chinese Language Processing*, pages 205-221. 1996.

[6] S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W.Furnas, and R.A. Harshman. "Indexing by latent semantic analysis." *Journal of the American Society for Information Science*, 41(6):391-407. 1990.

[7] Paolo Ferragina and Roberto Grossi. "The String B-tree: A new data structure for string search in external memory and its application." *Journal of ACM*, 46(2):236-280. 1999.

[8] William B.Frakes and Rick Kazman. Information Retrieval Data Structures Algorithm. Prentice Hall, Englewood Cliffs, New Jersey 0732. 1992.

[9] Marko Frobelink and Dunja Mladenic. "Turning yahoo into an automatic web-page classifier." *In Proceedings of the 13th European Conference on Aritficial Intelligence*, pages 473-474. 1998.

[10] Huang Sen-Yuan, Yi-Ling Chou, and Ja-Chen Lin. "Automatic classification for news written in Chinese." *Computer Processing of Oriental Languages*, 12(2):143-159. 1998.

[11] Leonard Kaufman and Peter J. Rousseeuw. "Finding Groups in Data Analysis : An Introduction to Cluster Analysis." John Wiley and Sons,Inc., New York. 1990.

[12] David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. "Training algorithms for linear text classifiers." *In Proceedings of the 19th Ann Int ACM SIFIR Conference on Research and Development in Information Retrieval* (SIGIR'96), pages 298-306. 1996.

[13] Lin Yih-Jeng, Ming-Shing Yu, Shyh-Yang Hwang, and Ming-Jer Wu. "A way to extract unknown words without dictionary from Chinese corpus and its applications." *In Research on Computational Linguistics Conference* (ROCLING XI), pages 217-226. 1998.

[14] Tom M. Mitchell. Machine Learning. The McGraw-Hill Companies, Inc. 1997 .

[15] Amitabh Kumar Singhal. "Term Weighting Revisited. " *PHD theses*, Cornell University. 1997.

[16] Tsay Jyh-Jong and Jing-Doo Wang. "Term selection with distributional clustering for Chinese text categorization using n-grams." *In Research on Computational Linguistics Conference XII*, pages 151-170. 1999.

[17] Tsay Jyh-Jong and Jing-Doo Wang. "Improving automatic Chinese text categorization by error correction." *In The Fifth International Workshop on Information Retrieval with Asian Languages*(IRAL2000), pages 1-8. 2000.

[18] Jyh-Jong Tsay and Jing-Doo Wang. "A scalable approach for Chinese term extraction." *In 2000 International Computer Sympoyium*(ICS2000), Taiwan, R.O.C, pages 246-253. 2000.

[19] R.C.Yang. *The Thesaurus of Daily Wordings*. Book-Spring Publishing Company, Taiwan. 1995.

[20] Yang Yiming and Xin Liu. "A re-examination of text categorization methods." *In Proceedings of the 22th Ann Int ACM SIFIR Conference on Research and Development in Information Retrieval*(SIGIR'99),pages 42-49. 1999.

[21] Yang Yiming and Jan O.Pedersen. "A comparative study on feature selection in text categorization." *In Proceedings of the Fourteenth International Conference on Machine Learning*(ICML'97), pages 412-420. 1997.

[22] Yang Yun-Yan. "A study of document auto-classification in mandarin Chinese." *In Research on Computational Linguistics Conference*(ROCLING VI), pages 217-233. 1993.

Computational Linguistics and Chinese Language Processing
Vol. 5, No. 2, August 2000, pp. 59-86

Japanese-Chinese Cross-Language Information Retrieval: An Interlingua Approach

Md. Maruf Hasan* and Yuji Matsumoto*

Abstract

Electronically available multilingual information can be divided into two major categories: (1) alphabetic language information (English-like alphabetic languages) and (2) ideographic language information (Chinese-like ideographic languages). The information available in non-English alphabetic languages as well as in ideographic languages (especially, in Japanese and Chinese) is growing at an incredibly high rate in recent years. Due to the ideographic nature of Japanese and Chinese, complicated with the existence of several encoding standards in use, efficient processing (representation, indexing, retrieval, etc.) of such information became a tedious task. In this paper, we propose a Han Character (Kanji) oriented Interlingua model of indexing and retrieving Japanese and Chinese information. We report the results of mono- and cross- language information retrieval on a Kanji space where documents and queries are represented in terms of Kanji oriented vectors. We also employ a dimensionality reduction technique to compute a *Kanji Conceptual Space* (KCS) from the initial Kanji space, which can facilitate conceptual retrieval of both mono- and cross- language information for these languages. Similar indexing approaches for multiple European languages through term association (e.g., latent semantic indexing) or through conceptual mapping (using lexical ontology such as, WordNet) are being intensively explored. The Interlingua approach investigated here with Japanese and Chinese languages, and the term (or concept) association model investigated with the European languages are similar; and these approaches can be easily integrated. Therefore, the proposed Interlingua model can pave the way for handling multilingual information access and retrieval efficiently and uniformly.

Keywords: Cross-language Information Retrieval; Multilingual Information Processing; Latent Semantic Indexing.

*Nara Institute of Science and Technology 8916-5, Takayama, Ikoma, Nara, 630-0101 Japan
E-Mail: {maruf-h, matsu}@is.aist-nara.ac.jp

1. Introduction

The amount of multilingual information available electronically has escalated in recent years. Lately, the information in non-English European languages and in Chinese, Japanese, Korean and Vietnamese (CJKV) is increasing at an incredibly high rate. Both Japanese and Chinese (also, Korean and Vietnamese, to some extent) are ideographic languages that use thousands of ideographic characters (also known as: Han characters, Kanji, Hanzi or Hanja) in writing. Managing the huge number of characters is no longer a problem in processing Japanese and Chinese language information. However, computer processing of these languages is complicated with the absence of word delimitation and the existence of several national and industrial encoding standards. Word delimitation (also called, *segmentation*) is an extra task to perform to process these languages, because in the written Japanese and Chinese texts, explicit boundaries between words are not available. Due to the existence of several encoding standards, it is also quite common to notice that most Internet search engines provide two different services to search for Chinese information in *Traditional Chinese* (commonly encoded in BIG-5 code) and the *Simplified Chinese* (GB code) form. Technically speaking, the tasks of processing and retrieval of traditional and simplified Chinese can be considered the tasks involving two distinct languages. Similarly, JIS, Shift-JIS and EUC, etc. are common Japanese encoding standards. The existence of several encoding standards complicates the information retrieval (IR) tasks for both Japanese and Chinese [24]. In this paper, we will formulate a unified framework to cope with the above-mentioned problems as well as to facilitate effective multilingual information retrieval.

Electronically available multilingual information can be divided into two major categories: (1) alphabetic language information (English-like alphabetic languages) and (2) ideographic language information (Chinese-like ideographic languages). Unicode, an increasingly popular encoding standard, defines uniform codes for almost all characters (both alphabetic and ideographic) of the world languages [43]. The common CJK ideographs section defined under Unicode is a superset of all ideographic Han characters used across the CJKV languages. This offers us an opportunity to represent Japanese and Chinese documents uniformly in Unicode. By doing so, we can also take advantage of Kanji to index and retrieve information across these languages. Nonetheless, the Kanji-derived semantic units or concepts can also be easily associated with the corresponding terms (stem, word or concept, etc.) of the alphabetic languages, and therefore, a universal multilingual IR framework can also be achieved.

The ubiquity of the Internet, the proliferation of electronic information and the emergence of globalization offer us the challenge of engineering sophisticated techniques to process multilingual and heterogeneous information efficiently. Therefore, in the recent years, the IR community put an exclusive focus on Cross-language Information Retrieval (CLIR) to address this new challenge [33]. CLIR investigates information indexing and retrieval issues across the languages. CLIR is a special case of Monolingual Information Retrieval (MLIR), and addresses the retrieval issues where queries and documents are given in different languages. If either the query or the document collection can be

effectively translated into the target language, the CLIR problems can be reduced to MLIR problems. The commonly used techniques for CLIR include three different approaches: (1) query translation, (2) document translation, and (3) combination of both query and document translation. However, given the fact that the quality of machine translation (MT) is still well below the desired level, CLIR often takes advantage of multilingual dictionaries, thesauri or word or sentence -aligned parallel corpora to circumvent MT. Also, there are CLIR approaches, which tend to bypass MT by making use of multilingual conceptual ontology [8] or multilingual term association [36]. Successful CLIR systems for European languages (English, French, Spanish and German, etc.) are demonstrated using conceptual mapping and term association techniques. In this paper, we investigate mono- and cross- language IR for Japanese and Chinese using Kanji mapping and semantic association of Kanji-derived concepts – a Kanji-based Interlingua CLIR for these ideographic languages.

Precisely speaking, we focus on the semantic information captured in Kanji and attempt to engineer the Kanji for effective mono- and cross- language information retrieval for Japanese and Chinese information. Unlike the characters (letters) of the non-ideographic languages (e.g., English, Arabic or Sanskrit), a single Kanji is capable of capturing significant semantic information within itself. However, single Kanji is ambiguous, and therefore, we attempted to index the Kanji through their explicit and implicit semantic contents. Despite our focus on these ideographic Asian languages, we have also included a discussion towards developing an Interlingua model of multilingual information processing, which is capable of handling other (non-ideographic) languages.

The organization of this paper is as follows. We briefly discuss the special issues of Japanese and Chinese information retrieval (IR) in Section 2. We include a detailed literature review of Japanese and Chinese IR in Section 3. In Section 4, we discuss several encoding standards of Japanese and Chinese texts, and their relationships with the Unicode. Our Kanji oriented retrieval experiments include four different indexing approaches: (1) single Kanji indexing, (2) single Kanji with Kanji bi-gram indexing, (3) single Kanji with correlated Kanji pair indexing (i.e., indexing the Kanji pairs that have high co-occurrence tendencies), and (4) Kanji based semantic indexing (i.e., by extracting latent Kanji concepts after applying the dimensionality reduction techniques). In Section 5, we introduce the vector space IR model in terms of Kanji vectors and Kanji-document matrix. The detail mathematical formalism of Kanji Co-occurrence Tendency (KCT) and Kanji Semantic Indexing (KSI) are outlined in Section 6. Finally, we discuss our mono- and cross- language IR experiments in Section 7, followed by a discussion and analysis in section 8. Throughout the entire article, whenever appropriate, we draw analogical comparisons between our approach and those of others to justify the formalism and the benefit of the proposed Interlingua model for multilingual information indexing and retrieval.

Readers who are familiar with the Japanese and Chinese language processing and vector space IR techniques, may skim through the introductory sections (Section 2, 3 and 4) and focus more on the later sections of this paper. Introductory sections are marked with asterisks.

2*. Special Issues in Japanese and Chinese Information Retrieval

In the following two Subsections, we will analyze the linguistic facets of the Japanese and Chinese languages, respectively, from an IR perspective.

2.1* Japanese IR

An investigation into popular Japanese full-text IR systems (NAMZU and FREYA, etc.) revealed that most Japanese IR systems mimic the IR systems for European languages [12, 28]. The indexing and retrieval usually involve with the four major steps: (1) *segmentation*: to locate word boundaries, (2) *morphological analysis*: to find the word-stems, (3) *representation and indexing of the stems*: e.g., to use inverted file or other data-structures to represent the document collection, and (4) *query-document similarity measurement*: to associate documents with queries using cosine or other similarity measures. The first two steps, segmentation and morphological analysis, are computationally expensive complex tasks, and these preprocessing steps result in a loss of syntactical cues from the documents and the queries. Given also the fact that vector space representation of documents and queries is a flat representation (known as a "bag of words"), which ignores contextual information of the original documents and the queries [37], it makes sense to represent the documents and queries only in terms of Kanji. We will investigate several ways of indexing Kanji (sometimes associated with further processing, such as Kanji mapping and correlation, etc.) straightforwardly to bypass computationally intensive segmentation and morphological analysis. It can be noted that, for Japanese-Chinese CLIR, such an approach can bring us an added advantage because the query or the document translation steps can be easily circumvented with Kanji association.

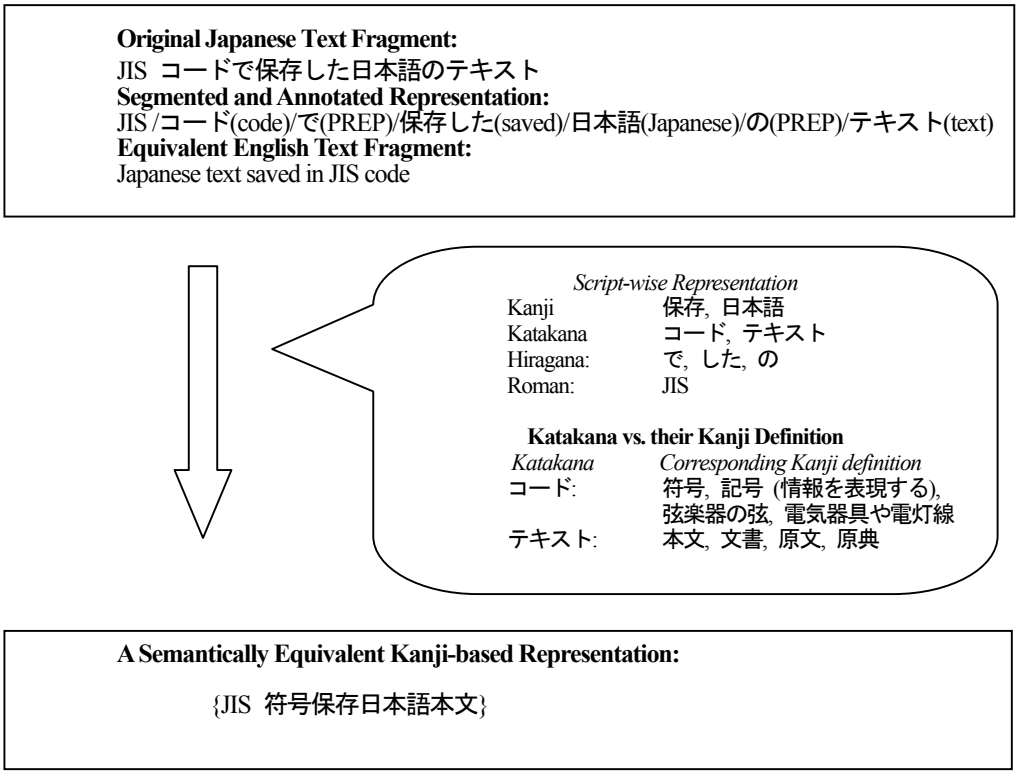


Figure 1 *A Japanese text fragment with four different scripts and its maximum likelihood mapping to corresponding Kanji*

Japanese texts are usually written using four different scripts: Kanji, Katakana, Hiragana and Roman alphabet [39]. The following example of a short Japanese text fragment (in Figure 1) shows the combination of all four scripts. In running Japanese texts, however, Kanji is more dominating than any other scripts. In writing, ideographic Kanji can be replaced with its Hiragana spelling. However, the Hiragana spellings are ambiguous. Kanji expresses the semantic more obviously than its Hiragana spelling; and probably because of this reason, despite a long ongoing debate on replacing Kanji completely with the Hiragana, Kanji still remains a major component in Japanese writing. Katakana is mostly used to transliterate loan words (except those borrowed from Chinese). However, the phonetic scarcity of Japanese, made the Katakana transliteration quite ambiguous. For example, the Katakana string, コード may represent different English words: code, cord and chord, etc. Moreover, Katakana transliteration is often inconsistent; the English word, "digital" can be written as デジタル or ディジタル, for instance. Most Katakana strings are content bearing terms and therefore, in Japanese IR, special care has always been taken to efficiently process the Katakana strings. In the ordinary Japanese-English dictionaries, a Katakana

entry appears with its original foreign word and a Japanese definition (often using Kanji). For example, the Katakana string, テキスト is defined with the relevant Kanji strings 文書, 原文, 原典 or 本文 as well as the English word, "text". In our Kanji based IR approach, we propose mapping the Katakana strings onto the relevant Kanji with the help of dictionary definitions. Hiragana strings are mostly shorter functional words or inflectional components. Short Hiragana strings usually play syntactical and morphological roles, and are usually ignored. However, continuous and long strings of Hiragana are potentially a replacement[1] for a rare and complicated Kanji (or a Kanji string), which, to some extent of accuracy, can also be replaced with the relevant Kanji. The Roman alphabet is usually mapped to the respective ASCII characters and indexed accordingly. For an effective Kanji-based information retrieval, we need to preprocess the Japanese documents and queries and represent them in terms of equivalent Kanji. Although sophisticated algorithms can be developed for Hiragana to Kanji and Katakana to Kanji mapping, we, for simplicity, use maximum likelihood based mapping strategy in this research. In Figure 1, we explain a Kanji based representation of the example text fragment using corresponding Kanji (the acronym, JIS, stands for the Japanese Industrial Standard).

It is worthy to note here that the complexity and the computational costs of such preprocessing are lesser than those of full segmentation and morphological analysis. Another justification for processing Japanese IR in this way is the usability of such indexing for CLIR without machine translation. Nonetheless, mapping a Katakana string to its original language is another feasible option for multilingual IR.

Although a single Kanji captures significant semantic information within itself, such information is highly ambiguous. Therefore, we also derived useful indexing information, such as, Kanji n-grams[2], correlated Kanji pairs and principal components, automatically from the initial Kanji based representation for effective indexing and retrieval.

2.2* Chinese IR

In comparison to Japanese IR, Chinese information retrieval is more straightforward because Chinese texts are mostly written homogeneously using only Kanji. However, like Japanese, Chinese text is also written without explicit word delimiters and therefore, segmentation must be performed to extract words from the string of Kanji for word or phrase level indexing. Since Chinese is a non-inflectional language, morphological analysis is not important. There are three major approaches to indexing Chinese text: (1) *single Kanji indexing,* (2) *Kanji n-gram indexing,* and (3) *word or phrase level indexing* (after segmentation). However, most practical systems incorporate more than one of the above indexing schemes

[1] In Japanese, materials written for the young people often include Hiragana strings to substitute rare Kanji.

[2] In this paper, we use the term, *n-gram* to refer to (n >1) cases. When n =1, we use the term, *single character indexing.*

for effective retrieval [32]. There are also reports on Chinese conceptual IR using a conceptual or semantic hierarchy [5].

The above discussion provides a fairly straightforward view of Chinese IR. However, due to the existence of several incompatible encoding standards, especially the Traditional (Big-5) and Simplified (GB) Chinese character encoding, Chinese IR is suffering a serious drawback. Moreover, from a technical standpoint, the Traditional Chinese IR and the Simplified Chinese IR can be viewed as two individual monolingual IR problems. This argument is further supported with the fact that most Internet search engines process information for simplified and traditional Chinese separately and offer two individual search interfaces for information retrieval in the specific encoding. We will discuss the coding related matters in Section 4 again after reviewing the CJK information retrieval literatures in the following section.

3*. A Collective Review of CJK Information Retrieval Related Work

In this section, we will present a brief review of CJK information retrieval. Although we will not report any experimental results on Korean IR in this paper, we have included a few Korean references in the appropriate contexts.

3.1* CJK Mono- and Cross- Language IR

Several approaches are investigated in CJK text indexing to address monolingual information retrieval (MLIR) - for example, (1) indexing words or phrases after segmentation and morphological analysis, (2) indexing *n*-gram ideographic characters, and (3) indexing single ideographic characters. From the potentially un-delimited sequence of characters, words must be extracted first. For word and phrase level indexing of the *inflectional* ideographic languages (e.g., Japanese and Korean), morphological analysis must also be performed. Sentences are segmented into words with the help of a dictionary using heuristic rules or machine learning techniques. Morphological analysis also needs intensive linguistic knowledge and computer processing. Segmentation and morphological analysis are complex tasks, and the accuracies of automatic segmentation and morphological analysis considerably vary across different domains. For the heterogeneous information sources, like the Internet, the accuracy of segmentation and morphological analysis perform more poorly than that of a particular domain. The computationally expensive word-based indexing of CJK texts, however, can contribute to better retrieval results when compared to the *n*-gram counterpart. Words and phrases are less ambiguous indexing units than the *n*-grams or the correlated *n*-grams, and therefore, can boost retrieval performance. Segmentation and morphological analysis related issues of Chinese, Japanese and Korean are intensively addressed elsewhere [40, 26, 18].

The n-gram (n >1) character-based indexing is computationally expensive as well. The number of indexing terms (n-grams) increases dramatically as n increases. Moreover, not all the n-grams are semantically meaningful words; therefore, smoothing and filtering heuristics must be employed to extract

lexically meaningful n-grams for effective retrieval of information. See [29, 30, 31, 4, 13, 19] for details. For European language CLIR, exciting experimental results are reported in [16] using n-gram character associations across English and French.

For Japanese and Chinese IR, indexing single character (Kanji) is straightforward and less demanding in terms of both space and time than those of n-gram or other indexing schemes. From a CLIR point of view, for single Kanji indexing, there is no need to (1) maintain a multilingual dictionary or thesaurus of words, (2) to extract words and morphemes, and (3) to employ machine learning and smoothing to prune trivial n-grams or to resolve ambiguity in word segmentation [21, 34, 22]. Moreover, for such a single Kanji based approach, there is no translation overhead for both queries and documents. This approach also eliminates some of the typical CLIR related problems discussed in [14].

Comparison of experimental results in monolingual IR using single character indexing, n-gram character indexing and (segmented) word indexing in Chinese information retrieval is reported in [19, 30, 31, 21]. For MLIR, n-gram and word -based approaches outperformed the single character based approach, at the cost of the extra time and space. Similar comparisons and conclusions for Japanese and Korean MLIR are made in [13] and [22], respectively.

Unlike MLIR, in cross-language information retrieval, a great deal of effort is allocated in maintaining the multilingual dictionary and thesaurus, and translating the queries and documents, and so on. In the CINDOR (**C**onceptual **IN**terlingua **DO**cument **R**etrieval) search from TextWise, LLC [8] uses a multilingual conceptual Interlingua approach for multilingual (English, French, Spanish and other European languages) information retrieval. Chen et al. [5] investigated conceptual CLIR of Chinese and English by mapping the WordNet Synsets to the concept hierarchy of a Chinese thesaurus. There are other approaches to CLIR where techniques like latent semantic indexing (LSI) are used to automatically establish associations between queries and documents independent of language differences [36]. Character tri-gram -based CLIR results [16] are also reported for English and French, which share a similar vocabulary. CLIR using Kanji association is also explored for ideographic languages like Japanese and Chinese [15]. However, no experiments have ever been conducted with a combination of ideographic and alphabetic languages through vocabulary association. This situation probably exists because of our practice of considering alphabetic and ideographic languages from different point of view. Such a practice should not be continued to foster truly multilingual information access and retrieval.

The authors sadly noticed that most of the CLIR research in recent years is focused on European languages. After the opening of the CLIR track in the TREC-6 conference [42], several reports have been published on cross-language information retrieval in European languages, and sometimes, European languages along with one of the Asian languages (e.g., Chinese-English, Japanese-English, etc.). TREC has not yet initiated any CLIR track that focuses on the Asian languages exclusively. In 1999, Pergamon published a special issue of the journal, Information Processing and Management focusing on Information Retrieval with Asian Languages [32]. Among the eight papers included in that special issue, only one

paper [19] addressed CLIR on multiple Asian language information retrieval (English, Japanese and Korean CLIR) using multilingual dictionaries and machine translation techniques (to translate both queries and documents within these languages). The recent initiative from the Asian Multimedia Forum (AMF) brought together three research institutes, NTT (Japan), KAIST (Korea) and KRDL (Singapore) to collaborate on CLIR research for CJK languages. However, their main focus is on using machine translation techniques for query and document translation, and thereby, integration of CJK information on the Internet to facilitate CLIR [2]. The series of conferences held in the name of IRAL (Information Retrieval with Asian Languages) addresses CLIR mostly in the traditional framework of query and document translation. For IRAL participants, the choice of languages still remains English and one of the Asian languages. We will investigate an Interlingua framework for Japanese-Chinese CLIR, which can easily be extendable to cover other languages.

3.2* Important Facts about Japanese and Chinese

Tan and Nagao [41] used Kanji correlation to align Japanese-Chinese parallel texts. According to them, the occurrence of common Kanji (in Japanese and Chinese language texts) sometimes is so prevalent that even a monolingual reader could perform a partial alignment of the bilingual texts. This fact can be further verified with the example in Figure 2. The common and visually similar Kanji appeared in news articles written in Chinese and Japanese provides enough cues to correlate the two reports (both describing a political crisis in the Korean Peninsula).

We would like to mention here that the named entities play an important role in IR and there is an intensive research focus on named entity extraction -related research. Not only is the extraction of named entities important, but the IR community must also work towards universal (Interlingual) representation and indexing of named entities for effective multilingual IR.

It should be noted that the pronunciations of the Kanji vary significantly across the CJK languages, but the visual appearances of the Kanji in written texts (across CJK language) have a certain level of similarity. The Unicode Kanji Information Dictionary [38] provides cross-references among all the unified CJK ideographs (encoded by the Unicode Consortium) across the CJK languages including the cross-reference between simplified and traditional forms of a Chinese character. As explained above, we may conclude that effective cross-language information retrieval by indexing and associating the non-trivial Kanji semantics holds promises. We can also bypass complicated segmentation or morphological analysis process using such an approach. At the same time, multilingual dictionaries and thesauri maintenance, as well as query and documents translations can also be avoided. In Section 7, we will report such experimental results.

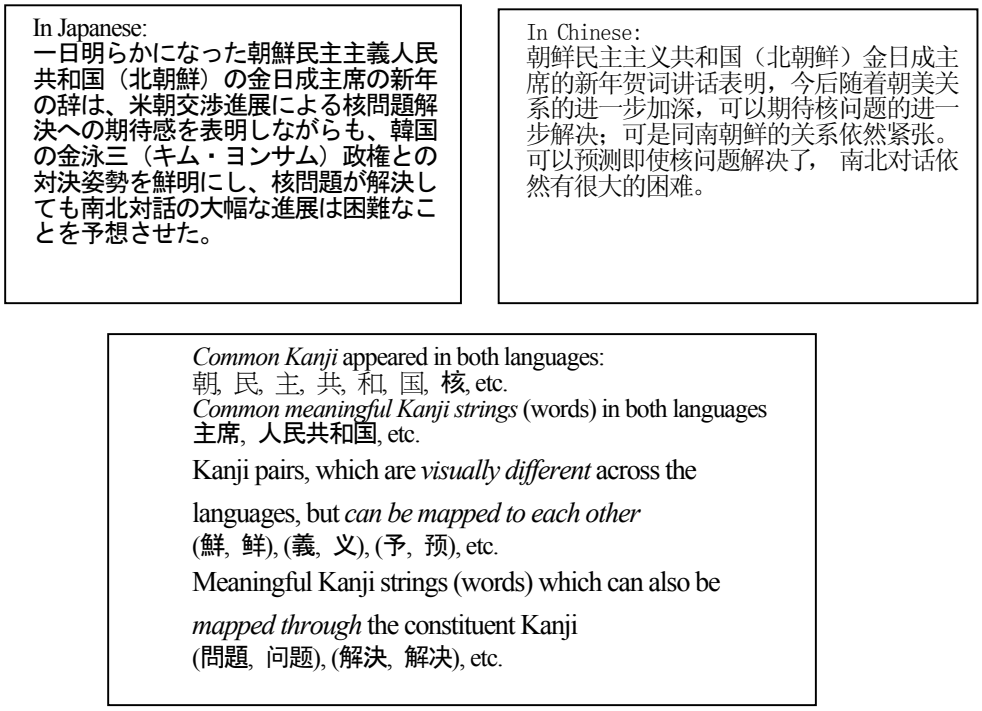


Figure 2 *Examples of common and similar kanji across Japanese and Chinese newswires*

4*. Japanese and Chinese Encoding Standards vs. the Unicode

Character encoding schemes of Japanese and Chinese have several variations: for example, Chinese encoding standards include two major standards- GB and BIG-5; and Japanese encoding standards include JIS, Shift-JIS and EUC, etc. A typical Internet search engine usually asks users to specify not only the language but also the encoding scheme (e.g., simplified (GB) or traditional Chinese (BIG-5)) while searching for information on the same language. For comprehensive details on different encoding standards, readers are referred to [24]. The number of Kanji encoded under a particular encoding scheme of a particular language also varies. Due to the huge difference in the number of Kanji encoded in simplified Chinese (GB) and in traditional Chinese (BIG-5), the retrieval of simplified and traditional Chinese are currently being processed as if they were two different languages. However, the problem with Japanese IR (due to the existence of different encoding standards) is not as severe as that of Chinese.

Unicode, a comprehensive coding scheme of the world languages, paves the way for an alternative and consistent representation of textual information.

Due to the growing acceptance and popularity of the Unicode [43] by the computer industry, we

have a common platform to investigate multiple languages in a unified framework. The Common CJK Ideograph section of the Unicode encoding scheme includes all characters encoded in each individual language and encoding scheme. Unicode version 3.0 assigns codes to 27,484 Kanji, a superset of characters encoded under the existing standards. Unicode makes it possible to represent documents uniformly across these languages. By representing Japanese and Chinese documents in Unicode and finding association through Kanji vocabulary it is possible to efficiently address the IR issues of these languages.

Figure 3 *Different ideographs represent the same concept, sword*

However, Unicode encoding is not a linguistically based encoding scheme; it is rather an initiative to cope with the variants of different local standards. A critical analysis of Unicode and a proposal of Multicode can be found in [27]. The Unicode standard avoids duplicate encoding of the same character; for example, the character 'a' is encoded only once although it is included in the alphabets of several western languages. However, for ideographic characters, such efforts failed to a certain extent due to the variation of typeface used under different situations and cultural settings. The ideographic characters in Figure 3, although they represent the same word (sword in English), are given six unique codes under the Unicode encoding scheme to satisfy the round-trip criteria[3], that is, to allow round-trip conversion between the source standard (in this case, JIS, which assigns 6 distinct codes) and the Unicode. The 27,484 Kanji encoded in Unicode, therefore, includes semantic redundancy in both single-language and multiple-language perspectives.

[3] A detail description of the *Unicode ideographic character unification rules* can be found in [43],

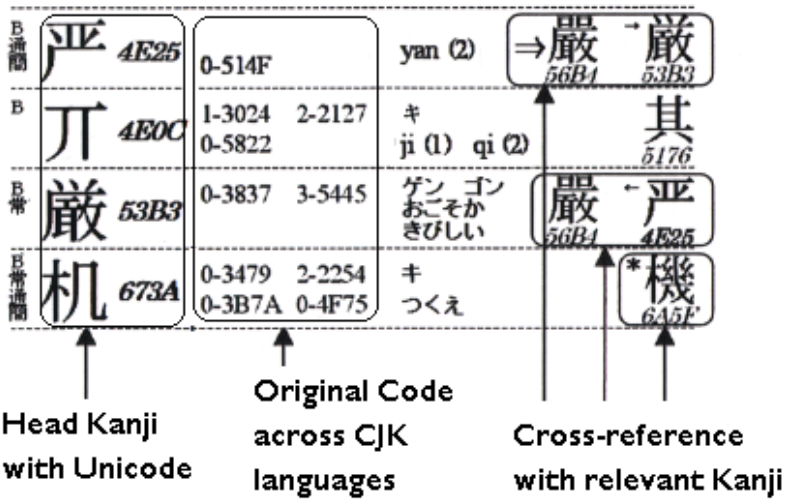


Figure 4 *Entries from the* Unicode Kanji Information Dictionary: *Kanji are annotated with cross-reference information within and across the CJK languages.*

In the unified CJK ideograph section, Unicode maintains redundancy to accommodate typographical or cultural compatibility because the design goal of Unicode is mainly to attain compatibility with the existing corporate and national encoding standards. In a Kanji-based CLIR approach, any such redundancy and multiplicity must be identified and resolved to achieve semantic uniformity and association within and across languages. Such tasks are less painstaking than the maintenance of multilingual dictionaries and thesauri of words and concepts. New lexical and ontological references, like the Unicode Kanji Information Dictionary [38] provides substantial co-reference information to assist such tasks. In our experiment, we use a table-lookup mapping approach to locate and associate the semantically related (but visually dissimilar) ideographs within and across CJK languages as a pre-processing task. Nonetheless, we are aware that there are cases where the same Kanji represents totally different concepts across the two languages in question. For the Kanji oriented CLIR, this phenomenon can somehow be considered as Kanji polysemy[4]. Such polysemy resolution can open up a new area of be further research (theoretically similar to word sense disambiguation problems).

5. Kanji-Document Matrix in Japanese and Chinese Information Retrieval

As justified above, in this research, we represent Japanese and Chinese documents uniformly using

[4] Kanji polysemy across the languages

Unicode. For Japanese information retrieval, since disregarding the Hiragana and Katakana could cause a partial loss in document and query semantics, we also perform further preprocessing by locating and replacing the lexically meaningful Hiragana and Katakana strings with the relevant Kanji (or Kanji string).

The next task is to index the Japanese and Chinese documents in terms of Kanji. Single Kanji indexing is the simplest among all types of indexing. Indexing Kanji n-grams (specially, bi-grams) is another possible alternative. Other correlation measures [10], used for calculating word co-occurrence of a language, are equally applicable to calculate Kanji correlation. We compute term frequencies (tf) and inverse document frequencies (idf) for single Kanji and Kanji bi-grams for indexing. We only consider bi-grams with medium frequency and exclude the most frequent and rare bi-grams based on empirically decided cut-offs. We also identify the correlated Kanji pairs based on the co-occurrence tendency measured using mutual information (c.f., Section 6.1). We only compute tf and idf of highly correlated Kanji pairs (decided by their mutual information measure) for indexing.

In the vector space IR model, a term-document matrix is computed from a collection of documents. Each column of the term-document matrix represents a document, and each row represents a term (e.g., a word, a phrase, a Kanji or a Kanji string). Each element of the matrix represents the weight of a term. For the simplest case, weights can be binary values, representing the presence or absence of a particular term in the document. The frequency of a term in a particular document normalized with the same term's inverse frequency with respect to the entire collection (generically known as, tf.idf) is also often used [37] as weights. A Kanji-document matrix is similar to a term-document matrix when we consider Kanji (one or more) as terms. We compute three different Kanji-document matrices using the tf.idf weighting scheme in three different ways: single Kanji (K_A for short), single Kanji with the Kanji Bi-grams (K_B), and single Kanji with the correlated Kanji pairs (K_C). These matrices are essentially the Kanji Space Representations of our document collection. Each column vector of the Kanji-document matrix is the Kanji Vector Representation of a particular document. Queries can also be represented as Kanji vectors. Relevance is computed by calculating the vector similarity between the query and the document collection.

5.1 Monolingual Kanji Conceptual Space (KCS)

The three different Kanji-document matrices introduced above are Kanji-vector representations of a document collection. *Kanji Conceptual Space* (KCS) is a conceptual representation of Kanji-concepts after projecting the original high dimensional Kanji vectors to a lower dimensional conceptual space. Although theoretically, we can compute three different KCSs from the three Kanji-document matrices, we will restrict us in computing only one KCS from the single Kanji based Kanji-document matrix (using the *log-entropy* weighting scheme). This restriction is due to the constrain that with a small collection of documents, if we include the Kanji bi-grams or the Kanji correlated pairs as terms, the total number of terms (including single Kanji) exceeds the total number of documents in the collection. This situation violates the assumption behind SVD as explained in Section 6.2. Moreover, using SVD to reduce the

dimension of a heterogeneous space (estimated using single Kanji along with the correlated Kanji pairs or Kanji bi-grams estimated from a small collection of documents) into a reduced space may not achieve a proper conceptual mapping.

We used the log-entropy weighting, (log (tf + 1) . entropy) as described in [9], where tf and the entropy are the term frequency and entropy of a Kanji. We chose the log-entropy weighting scheme for three reasons, (1) it is faster to compute, (2) other LSI-based experiments intensively use log-entropy measure, and (3) we verified (using non-parametric Wilcoxon matched-pair sign test) that the difference between the tf.idf weighting scheme and the log-entropy weighting for single Kanji indexing is insignificant[5]. Other weighting schemes that incorporate a local and a global factor (such as, tf.idf variants) may also be applicable.

Latent Semantic Indexing (LSI), a well-known vector space model of IR, is capable of performing conceptual information retrieval. LSI uses the singular value decomposition (SVD) technique to reduce the rank of the original term-document matrix. Theoretically, SVD, a principal component analysis technique, performs a term-to-concept mapping and therefore, conceptual indexing and retrieval is made possible [7]. Considering the computational overhead of SVD, we chose the log-entropy weighting scheme of single Kanji, which can be computed faster [9], and computed single-Kanji based Kanji-document matrices for both Japanese and Chinese documents. By applying SVD to these Kanji-document matrices, we can derive conceptual representations of a text object (a document or a query) on the Kanji conceptual space in the respective language. For convenience, we will refer to this dimensionality reduction based approach as K_D.

5.2 Cross-language Kanji Conceptual Space: An Interlingua Representation

For CLIR experiments of European languages, Rehder et al. [36] experimented with English-French-German CLIR using a multilingual parallel corpus of these languages. They decomposed the multilingual term-document matrix using SVD to find associations of words (e.g., vocabulary mapping) among these languages. Interlingual conceptual representation of a document or a query can be computed from decomposed multilingual term-document representation since such a representation captures significant information about cross-language vocabulary mapping [11].

As described in Section 5.1, the rank-reduced Kanji-document representation for Japanese (or Chinese) documents can be used to represent a Japanese (or Chinese) document or a query in the Japanese (or Chinese) Kanji conceptual space. Similarly, with a collection of properly aligned Japanese and Chinese parallel documents, it is possible to compute a reduced rank Kanji-document matrix on the unified Kanji space, where each column represents an aligned pair of bilingual documents (in terms of all the Kanji that appeared in the Japanese documents and in its corresponding Chinese document). A moderate size of such

[5] Wilcoxon tests can measure whether the difference in retrieval results between two experiments is significant [17].

a parallel corpus can capture the bilingual Kanji Conceptual Space across the two languages. Decomposing this bilingual Kanji-document matrix into a conceptual space theoretically enables us to compute an Interlingual Kanji Conceptual Space. Both mono- and cross- language information retrieval can be efficiently performed on this unified KCS. We discussed the mathematical formalism of such an approach in Section 6.2.

Since we did not find any suitable parallel corpora of Japanese and Chinese documents, we used a commercial MT system to translate the Japanese document collection into Chinese, and the Chinese collection into Japanese. Assuming that the quality of the machine translation has a trivial impact (since we are only interested in finding the Kanji association), the original documents and their translations provide us an alternative opportunity to roughly estimate the bilingual KCS. First, we computed a bilingual Kanji-document matrix using the log-entropy of each Kanji. By reducing the rank of this bilingual matrix using SVD, we can derive an Interlingual representation of our bilingual document collection on the unified KCS.

We want to conclude this section by pointing out that such an Interlingua representation, although derived from ideographic Kanji, is easily extendable to non-ideographic languages because of the flexibility of representation in the vector space model. Rather than associating Kanji with the word-stems of each alphabetic language, we consider mapping the words to their respective word-roots. Since the vocabulary of many European languages can be mapped back to their original roots (Latin, and Greek, etc.), word-roots associated with the Kanji can provide a multilingual conceptual space for effective representation and retrieval of multilingual information.

6. Kanji Co-occurrence Tendency and Kanji Semantic Indexing

6.1 Calculation of Kanji Co-occurrence Tendency (KCT)

Counting the weights for a single Kanji and that of bi-grams are straightforward and we skip the details for brevity. Here, we will define the *Kanji Co-occurrence Tendency* (KCT) and explain how we computed KCT and choose the correlated Kanji pair.

Mutual Information (MI) is one of the metrics that can be used for calculating the significance of word co-occurrence associations [6, 25]. We extended the idea to estimate KCT. The mutual information MI between two events x and y is defined as follows:

$$MI_2(x, y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where $p(x,y)$ is the joint probability that two events, x and y co-occur, and $p(x)$ and $p(y)$ are the probabilities that event x or y occurs independently.

MI can be applied to the Kanji in the documents and can be used to calculate the correlation between those Kanji. To calculate the co-occurrence tendency for two Kanji, k_1 and k_2, we define n_{ij} (i, j = 1,2) in a 2-by-2 table shown below. In the table, n_{11} indicates the number of times two Kanji, k_1 and k_2 co-occur within a text window, n_{12} indicates the number of times k_1 occurs, but k_2 does not occur within a text window, and so on.

Table 1. *Auxiliary Table for defining Kanji Co-occurrence Tendency*

	$\mathbf{k_2}$	$\sim \mathbf{k_2}$
$\mathbf{k_1}$	n_{11}	n_{12}
$\sim \mathbf{k_1}$	n_{21}	n_{22}

$n_{i.}$, $n_{.j}$ and N are defined as follows:

$$n_{i.} = n_{i1} + n_{i2}$$

$$n_{.j} = n_{1j} + n_{2j}$$

$$N = \sum_{i,j} n_{ij}$$

That is, $n_{i.}$ indicates the number of times k_1 occurs (i = 1) or does not occur (i = 2) regardless of the occurrence of k_2, and N indicates the total number of co-occurrence windows in the corpus.

The co-occurrence tendency of a Kanji pair k_1 and k_2 in a corpus, KCT_{MI2} is defined as follows:

$$KCT_{MI2}(k_1, k_2) = \log_2 \frac{\frac{n_{11}}{N}}{\frac{n_{.1}}{N}\frac{n_{1.}}{N}} \quad (1)$$

Note that we use one entire document as the window of co-occurrence instead of a fixed number of words. Usually, co-occurrences are measured between two Kanji mainly because of computational and storage costs. We can use co-occurrence frequencies among any n Kanji. n Kanji co-occurrence tendency KCT_{MIn} among Kanji, k_1 , k_2 ...k_n is defined, as an extension of KCT_{MI2}, as follows:

$$KCT_{MIn}(k_1, k_{2,\ldots}k_n) = \frac{1}{n-1}\log_2 \frac{\frac{f(k_1,k_{2,\ldots}k_n)}{N}}{\frac{f(k_1)}{N}\frac{f(k_2)}{N}\cdots\frac{f(k_n)}{N}}$$

where N is the total number of documents in a document collection (corpus), f(k) is the number of documents where the Kanji k occurs, and f(k_1, k_2, ...k_n) is the number of documents where all Kanji, k_1 , k_2, ... k_n appear. Note that MI is essentially a measure between two events, so this is an ad hoc extension only for the purpose of calculating n-Kanji co-occurrence tendency. We will only calculate MI for Kanji pairs to select highly correlated Kanji pairs for indexing.

We would like to add that other co-occurrence tendency measures, for example, dice co-efficient, log likelihood ratio and Chi-square test [10, 25] are also applicable to calculate KCT-like measures.

6.2 Latent Semantic Indexing (LSI) and Kanji Semantic Indexing (KSI)

Like Latent Semantic Indexing (LSI), KSI begins with a collection of *m* documents containing *n* unique Kanji (i.e., terms) and KSI forms an *n x m* sparse matrix *A*, with A_{ij} containing a value related to the number of times Kanji *i* appears in document *j*. Various weighting schemes can be applied to the raw occurrence counts. In this work, we used log-entropy weighting.

Once the Kanji-document matrix *A* has been created, KSI computes the similarity between two text objects (a query and a document, for example) as follows. First, a text object *q* is represented by an *n x 1* vector, much like a column of the *A* matrix and with the same sorts of term weighting applied. Next, the similarity between text objects, q_1 and q_2 can be computed, typically by cosine scoring in the vector space model. This similarity can be represented as, $\mathbf{sim}\ (q_1, q_2) = q_1^T \quad q_2 / \sqrt{(q_1^T q_1 \,.\, q_2^T q_2)}$.

A mathematically useful way of viewing the process of computing text-object similarity scores in the vector space model is: (1) Each of the *n* Kanji in the collection has a vector representation. Specifically speaking, Kanji *i* is an *n x 1* vector of zeros with a *1* in component *i*; (2) The representation of a text object, *q* is the weighted sum of the Kanji vectors of all the Kanji that appear in the text object. Thus, the similarity between text objects, q_1 and q_2 is:

$$\mathbf{sim}\ (I_n q_1\ , I_n q_2\ ) \qquad (2)$$

where I_n is the *n x n* identity matrix. Here, I_n plays the role of a *vector lexicon*, in the sense that it assigns each Kanji a vector definition. Of course, pre-multiplying by the identity matrix in the Eq. 2 does not change the comparison in any way. On the other hand, by using other vector lexicons, we can substantially change the way similarities are computed. In addition, the only role played by the Kanji-document matrix *A* in the vector space model is in the computation of weighting factors for the components (i.e., Kanji or terms) of text objects.

KSI, like LSI, is a vector space IR formalism. KSI begins with the formation of the Kanji-document matrix *A*. Then, the *A* matrix is analyzed using singular value decomposition (SVD) to extract structure concerning document-document and Kanji-Kanji correlations. Mathematically, an SVD of *A* can be written as,

$$A = \mathrm{U}(A)\, \Sigma(A)\, \mathrm{V}(A)^T \qquad (3)$$

where $\mathrm{U}(A)$ is an *n x n* matrix such that $\mathrm{U}(A)^T\, \mathrm{U}(A) = I_n$, $\Sigma(A)$ is an *n x n* diagonal matrix of singular values, and $\mathrm{V}(A)$ is an *n x m* matrix such that $\mathrm{V}(A)^T\, \mathrm{V}(A) = I_m$. This assumes for simplicity of exposition that *A* has fewer terms than documents, $n < m$.

This SVD analysis can be used to construct lower rank approximations of *A*, and this is how it is

typically used in the context of LSI. Reducing the rank of the approximation results in a synonym collapsing effect in practice. Such a reduction also lessens the total amount of processing and storage overheads associated with preprocessing and retrieval. We use A_k to denote the components of the k-dimensional SVD, and the rank-k reconstruction of A as follows:

$$A_k = \mathrm{U_k}(A)\, \Sigma_k(A)\, \mathrm{V_k(A)}^{\mathrm{T}} \qquad (4)$$

The $\mathrm{U_k}(A)$ matrix in Eq. 4 can be used as an alternative vector lexicon to the I_n in Eq. 2 in that it assigns a vector representation to every Kanji in the Kanji-document matrix A. Thus, in KSI, the k-dimensional similarity between text object q_1 and text object q_2 in the context of A is,

$$\mathbf{sim}\,(\mathrm{U_k(A)}^{\mathrm{T}}\, q_1\,,\, \mathrm{U_k}(A)^{\mathrm{T}}\, q_2) \qquad (5)$$

In the LSI literature [7, 9, 11, 3, 36] justifications for the use of the matrix of left singular vectors $\mathrm{U_k(A)}$ as a vector lexicon are intensively studied.

Cross-language LSI (CL-LSI)

The techniques of monolingual LSI can be extended easily to the cross-language case simply by using a different notion of the term-document matrix. For concreteness, let E be a term-document matrix of m English documents and n^E English terms, and let F be a term-document matrix of m semantically equivalent French documents and n^F French terms. These documents are aligned pair-wise, in the sense that document $1 \le i \le m$ in the English collection is directly related to document i in the French collection. The multi-language term-document matrix M can be written as follows:

$$M = \begin{bmatrix} E \\ F \end{bmatrix}$$

M is an $(n^E + n^F)$ *x* m matrix in which column i is a vector representing the English and French terms appeared in the *union* of document i written in both languages. Cross-language LSI (CL-LSI) begins with the matrix M and performs an SVD,

$$M = \begin{bmatrix} U_k^E(M) \\ U_k^F(M) \end{bmatrix} \Sigma_k(M)\; V_k(M)$$

where $\mathrm{U_k^E}\,(M)$ and $\mathrm{U_k^F}\,(M)$ are k-dimensional vector-lexicons for English and French, respectively. Empirically, similar English and French words are given similar definitions, so this vector lexicon can be used for cross-language retrieval. In particular, consider an English text object $\mathrm{q^E}$ and a French text object q^F. They can be compared using the obvious generalization of Eq. 5,

$$\mathbf{sim}\,(\mathrm{U_k^E}(M)^{\mathrm{T}}\, \mathrm{q_E}\,,\, \mathrm{U_k^F}(M)^{\mathrm{T}}\, \mathrm{q_F}) \qquad (6)$$

In our experiments, we chose Japanese and Chinese, and represented the Japanese and Chinese document and query using Unicode, where each Kanji is equivalent to a term. Common or semantically

similar Kanji are considered as cross-language homonyms. Differences in Kanji usage across the languages are captured through SVD decomposition. We refer to such representation and indexing as *cross-language Kanji Semantic Indexing*.

7. Experimental Setups and Results

7.1. Document Collection and Queries

The most popular IR Test Collections in Japan are the NTCIR collection and the BMIR-J2 collection. The NTCIR collection consists of a collection of 330,000 English and Japanese scientific articles. Half of the collection (187,081 documents) consists of bilingually aligned English-Japanese document pairs. This collection was not suitable for our experiments since we want to experiment with Chinese as well. The BMIR-J2 Test Collection is a Japanese text collection of 5,080 newspaper articles chosen from the Mainichi newspaper. However, due to the participants' interests, this collection is restricted to the Engineering and Economics domains only. We could use this collection along with a set of corresponding Chinese document collection (not strictly parallel, but comparable counterparts). However, locating corresponding Chinese articles in the Engineering and Economics domains using Search Engines or Internet Robots is a tedious task.

Another international test collection, the TREC test collection includes English-Chinese and English-Japanese parallel corpora but no test collection of Japanese-Chinese parallel documents is yet available. Since we want to experiment with Japanese and Chinese IR and since there is no Japanese-Chinese bilingual test collection available so far, we took the initiative to prepare a bilingual test collection for our use. Nevertheless, we adhered to the TREC and NTCIR guidelines as strictly as possible. For the ease of locating corresponding Chinese documents, we restricted ourselves to current international affairs. We will explain the details of our collection preparation procedures below.

Mainichi newspaper is publishing their newswire archive on the CD-ROM on a yearly basis since early 90s. First, we used full-text search engines to index the most recent archives of the Mainichi Shimbun newspaper articles. Initially, we constructed 50 initial queries by selectively examining the collection. We used the freely available full-text search tools (NAMAZU and FREYA) to retrieve documents (likely to include both relevant and non-relevant) from a selected portion of the most recent Mainichi Newspaper archives [1995-1999]. For each query, we retained the top 20 documents retrieved by each search engine. By merging a few hundred documents from each search engine, we obtained a collection of about 1,600 documents. We carefully investigated the 224 articles retrieved by both the search tools against our 50 initial queries. Finally, we settled on a set of 33 *revised* queries and 1,000 news articles. After deciding on a set of documents and queries, we re-indexed this collection of 1,000 articles

and used the polling method[6] to prepare the query-relevance matrix. Our polling process is highly approximated because we used the output results of only two systems validated by a single human evaluator. The accuracy of the query-relevance matrix can be further improved by employing more human evaluators.

After deciding on the set of Japanese queries and documents, we translated the queries into Chinese and used Internet search engines to retrieve and choose a collection of 1,000 Chinese documents using advanced search features provided by AltaVista [1]. Because our Japanese document collection mostly consists of articles about current and International affairs in recent years, it was easier to locate similar Chinese articles on the Internet. Moreover, the AltaVista search engine facilitates restricted-search within a particular domain. By restricting us to news sites, we could quickly extract a collection of Chinese documents comparable to the Japanese document collection. We again used the polling strategy to compute a query-relevance matrix for this collection with the help of two search tools and one human evaluator.

7.2. Retrieval Results for Monolingual IR

We convert the entire bilingual collections (including queries) into Unicode. Necessary preprocessing (e.g., Kanji mapping) of the document collection is also done prior to indexing. We use a modified version of the publicly available *mg* System [45] developed as part of the New Zealand digital library (NZDL) project to index our document collection in 3 different ways:

1. K_A, Single Kanji indexing
2. K_B, Kanji bi-gram indexing, and
3. K_C, Correlated Kanji pair indexing

We also use the LSI++ [23] package for singular value decomposition of the Kanji-document matrices (computed with the log-entropy weighting of single Kanji), and investigate the latent Kanji semantic retrieval.

4. K_D, reducing the dimension of the Kanji-document matrix using SVD

We use the TREC evaluation scripts (TREC-EVAL) to compute the non-interpolated average precision for the 33 queries. The monolingual retrieval results are listed in Table 2, for both Chinese and Japanese information retrieval.

In Table 2, the average precision is very low for single character indexing and bi-gram indexing. In general, the average precision for our Japanese and Chinese monolingual IR are far below the average precision level achieved by the TREC, NTCIR and BMIR-J2 participants. The poor average precision is

[6] The polling method is described in [44] and used by TREC, NTCIR and other test administering authorities.

because of the fact that we do not incorporate classical IR enhancement mechanisms such as, query expansion and relevance feedback. The single character indexing and bi-gram indexing approaches are the simplest approaches compared to the extensive linguistic and computational techniques employed by the NTCIR and BMIR-J2 participants. Since our focus is to investigate the effectiveness of a Kanji-based representation, we refrain from enhancing the retrieval procedures through complex mechanisms so that we can investigate the true effect of such representation and indexing. For the same reason, direct comparison of the results of our experiments with those of others is not possible.

Table 2. *Average precision for Japanese and Chinese Monolingual IR*

Indexing Method	**Non-interpolated Average Precision**	
	1,000 Japanese documents & 33 Queries	1,000 Chinese documents & 33 Queries
K_A (Single Kanji)	.1435	.1838
K_B (+Kanji Bi-gram)	.1626	.2253
K_C (+co-occurrence, MI)	.1757	.2482
K_{D30} (log +SVD, k=30)	.1505	.2037
K_{D100} (log +SVD, k = 100)	.1870	.2528

By incorporating extra computation efforts for the singular value computation (i.e., without linguistic analysis), we achieve a significant boost in the average precision for the case of a relatively larger value of k (k = 100 to 300). A single Kanji itself is ambiguous but when groups of Kanji are mapped into a reasonable number of concepts, the latent concept of the query and the documents matches efficiently.

For a smaller value of k (k = 30), we perform some error analysis and discover that the performance degradation is severe for the short-queries, for which a large number of non-relevant documents are also retrieved with high ranking. Kanji semantic indexing (KSI) may perform better with a proper mechanism of query expansion for the short queries. Another potential problem with KSI is that the parameter, k, may have to be adjusted depending on the nature of the collection. For our case, the retrieval results with k=30 and 100 are chosen empirically. For k value within 100 to 300, we obtain stable retrieval performance.

The overall Chinese retrieval results are better than those for Japanese, perhaps due to the homogeneous Chinese scripts. For Japanese IR, a plausible way of improving the retrieval efficiency is to put more effort in Katakana disambiguation and Hiragana to Kanji conversion.

7.3. Retrieval Results for Cross-language IR

In this section, we discuss three different retrieval results. First, we use Japanese queries to retrieve Chinese documents. Second, we use Chinese queries to perform retrieval from the Japanese collection. Note that in both experiments neither query translation nor document translation is performed. Documents are retrieved in terms of Kanji correspondence. The third experiment is performed under a special condition where

commercial MT system is used to translate the Japanese and Chinese documents and a pseudo-query expansion mechanism (described later in this section) is employed.

From the CLIR results shown in the 2nd and 3rd columns of the Table below (Table 3), it can be concluded that the CLIR results using only Kanji mapping and associations are not prohibitive for the K_A and K_B, where single Kanji and the Kanji bi-grams are the basis of indexing and retrieval. However, the retrieval results are very promising for the K_C and K_D, where Kanji correlation and Kanji association are the basis of indexing and retrieval. Theoretically, it can be said that the retrieval results can be further improved if the Kanji correlation and the Kanji association are estimated from a large collection of documents.

The 4th column of Table 3 shows the retrieval result under a special situation. We use a commercial MT system (Chinese-Japanese/Japanese-Chinese) from Kodensha [20] to translate the Chinese document collection into Japanese, and vice versa. This MT system uses a basic bilingual dictionary of 120,000 Japanese-Chinese and 220,000 Chinese-Japanese entries. After the translation, we append the translated documents with the respective originals. In this way, we have a bilingual document collection of 2,000 documents. Since we are only considering Kanji and Kanji derived information in our indexing process, we assume that the quality of the machine translation (in terms of readability, syntax, etc.) has trivial effects. We also merge the corresponding queries in Chinese and Japanese to obtain the pseudo- *query translation* and pseudo- *query expansion* effects. Please note that the Kanji semantic retrieval, K_D is based on log-entropy weights and the other three approaches are based on the *tf.idf* weighting scheme.

Table 3. *Average Precision for Japanese and Chinese Cross-Language IR*

Indexing Method	**Non-interpolated Average Precision**		
	1,000 Japanese documents & 33 Chinese queries	1,000 Chinese documents with 33 Japanese queries	2,000 bilingual documents with 33 merged queries
K_A (Single Kanji)	.1033	.1398	.1241
K_B (+Kanji Bi-gram)	.1244	.1569	.1348
K_C (+co-occurrence, MI)	.1656	.1972	.2024
K_{D30} (+SVD, k = 30)	.1547	.1780	.2326
K_{D100} (+SVD, k = 100)	.1622	.2016	.2537

The non-interpolated average precisions of document retrieval using this approach with 33 merged queries and 2000 bilingual documents are listed in the 4th column of Table 3. The bi-gram based method (K_B) suffers from low precision. This is possibly due to MT-related errors. We assume that a properly aligned Japanese-Chinese parallel or comparable corpus may boost the bi-gram based retrieval results as well as the single Kanji based retrieval results (K_A). The co-occurrence based method (K_C) and Kanji association based method (K_D) perform better with the translated bilingual documents and merged queries.

From the above scenario of Kanji-based monolingual and cross-language information retrieval of Japanese and Chinese, we can safely conclude that by estimating Kanji correlation and Kanji association

from a large parallel (or comparable) corpus, it is possible to formulate effective Japanese-Chinese mono- and cross- language IR.

8. Discussions

In this paper, we explored one of the few possibilities of cross-language IR research with Asian languages. We experimented with Japanese and Chinese, where Kanji play an important semantic role; and we demonstrated that mono- and cross- language IR can be performed effectively through Kanji associations. In our experiments, we deliberately used Kanji and Kanji-derived semantics to address Japanese and Chinese IR (including CLIR). It is worthy to mention here that we do not advocate abandoning linguistic enhancements (e.g., segmentation, morphological analysis, etc.) and classical IR enhancements (e.g., query expansion, relevance feedback, etc.) techniques for IR tasks. Our exclusive attention to Kanji in our experiments is to identify the role of Kanji semantics in IR and CLIR. This approach can easily accommodate other linguistic and IR enhancements, and with such enhancements, the proposed approach will eventually give birth to practical CLIR systems.

Several types of ambiguities with Kanji usage across the Japanese and Chinese languages [15] exist. Such ambiguities contribute highly to the lower precision in single Kanji oriented indexing and retrieval. For bi-gram based indexing, we cannot conclude anything with high confidence due to the small document collections used in our experiments because of the data sparseness problem. However, the average precision of mono- and cross-language IR with KCT and with SVD shows that Kanji based indexing and retrieval of these ideographic languages is effective. Unlike other IR research reports where the IR task is comprehensively addressed, our experiments involves with only a straightforward hypotheses. Because of our exclusive focus on Kanji semantics and Japanese-Chinese language pair, we could not make direct comparison of our experimental results with those of the others'.

For Japanese-Chinese CLIR, this is one of the very first reports. The indexing methods we tried inherently bypass the complicated segmentation and morphological analysis phases, which would otherwise be necessary. Nonetheless, incorporating such linguistic analyses with the proposed approach will certainly improve the retrieval results. We are also aware that query expansion and relevance feedback -based enhancements can also be easily incorporated with the proposed Interlingua framework since this framework uses a flexible vector space representation. Moreover, Kanji association makes cross-language IR simpler than the traditional MT-based approach. We mapped Katakana and Hiragana strings to their relevant Kanji using an ad-hoc approach. During the error analysis, we noticed that such an approximated mapping significantly contributed to erroneous retrieval. Accurate mapping of Katakana and Hiragana strings to Kanji can further boost the retrieval effectiveness.

Since words in the non-ideographic languages can also be mapped to their original roots, the proposed Kanji-based Interlingua framework can be extended to deal with multilingual information indexing and retrieval of any combination of languages as far as parallel (or comparable) corpora and

sufficient computing power are available. Effective processing of multilingual heterogeneous information in this Internet age is inevitable. Revolution in storage capacity, abundance in computing power and invention of sophisticated mathematical methods for dimensionality reduction and projection (e.g., [23]) may present us with a better opportunity to integrate alphabetic and ideographic languages equally effectively under a uniform Interlingua representation. Indexing and retrieving heterogeneous multilingual information in a unified manner will therefore be possible. Traditional lexical (or Boolean) IR techniques will continuously be less effective as the number of documents grows. Multilingual IR is inevitable because of the global connectivity and the proliferation of electronic information. Automated and conceptual IR techniques will therefore dominate the future IR research.

Acknowledgements

We would like to extend our thanks to the people involved with NZDL, NAMAZU and FREYA projects. Their tools helped us to speed up our research. Thanks to Dr. Akira Maeda for allowing us to use his correlation calculation tool and Dr. Michael Berry for the LSI++ and SVDPACK packages. We thank the anonymous reviewers for their valuable comments.

References

[1] ALTAVISTA, "Altavista Advanced Search Tutorial", http://doc.altavista.com/adv_search/ast_toc.html

[2] AMF, "Cross-Language Information Retrieval at AMF - For Overcoming the Language Barrier in the Use of Internet", Asian Multimedia Forum, http://www.ntt.co.jp/news/news99e/9902/990224a.html

[3] M. Berry and P. Young, "Using Latent Semantic Indexing for Multi-Language Information Retrieval", *Computers and the Humanities*, 29(6), pp. 413-429, 1995.

[4] A. Chen, J. He, L. Xu, F.C. Gey and J. Meggs, "Chinese Text Retrieval Without Using a Dictionary", *In Proceedings of the Conference on Research and Development in Information Rertrieval*, ACM SIGIR-97, pp. 42-49, 1997.

[5] H.H. Chen, C.C. Lin and W.C. Lin, "Construction of a Chinese-English WordNet and its application to CLIR", *In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, IRAL-2000, pp. 189-196, 2000.

[6] K.W. Church and P. Hanks, "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics*, Vol. 16(1), pp. 22-29, 1990.

[7] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas and R.A. Harshman, "Indexing by latent semantic analysis," *In Journal of the American Society for Information Science*, 41(6), 391-407, 1990.

[8] A. Diekema, F. Oroumchian, P. Sheridan and E.D. Liddy, "TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French", *In Proceedings of the Seventh Text REtrieval Conference* (TREC-7), pp. 169-180, 1999. http://www.cindorsearch.com/

[9] S.T. Dumais, "Improving the Retrieval of Information from External Sources", Behavior Research Methods, Instruments, & Computers, Vol. 23, pp. 229-236, 1991.

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

[11] D.A. Evans, S.K. Handerson, I.A. Monarch, J. Pereiro, L. Delon and W.R. Hersh, "Mapping Vocabularies Using Latent Semantics", In G. Grefenstette Ed., *Cross-Language Information Retrieval*, Kluwer Academic Publisher, pp. 63-80, 1998.

[12] FREYA, *Full-text Retrieval Engine for Your Archive,* http://www.ingrid.org/ja/project/freya/ (in Japanese).

[13] H. Fujii and W.B. Croft, "A comparison of Indexing for Japanese Text Retrieval", *In Proceedings of the ACM SIGIR-93*, pp. 237-246, 1993.

[14] G. Grefenstette, "The Problem of Cross-Language Information Retrieval", In G. Grefenstette Ed., Cross-Language Information Retrieval, Kluwer Academic Publisher, pp. 1-10, 1998.

[15] M.M. Hasan and Y. Matsumoto, "Chinese-Japanese Cross Language Information Retrieval: A Han Character Based Approach", *In Proceedings of the SIGLEX Workshop on Word Senses and Multi-linguality*, pp.19-26, ACL-2000, Hong Kong, 2000.

[16] J. Hochberg and D. Nix, "Vector Mapping CLIR with Character Trigrams", *Los Alamos National Laboratory (LANL) CLIR Project Notebook Paper*, 1999. http://citeseer.nj.nec.com/134994.html

[17] D. Hull, "Using Statistical Testing in the Evaluation of Retrieval Experiments", *In Proceedings of the ACM SIGIR-93*, pp.329-338, 1993.

[18] D.B. Kim, K.S. Choi and K.H. Lee, "A Computational Model of Korean Morphological Analysis: A Prediction-based Approach", *Journal of East Asian Linguistics*, Vol. 5(2), pp. 183-215, 1996.

[19] T. Kim, C.M. Sim, S. Yuh, H. Jung, Y.K. Kim, S.K. Choi, D.I. Park and K.S. Choi, "FromTo-CLIR™: web-based natural language interface for cross-language information retrieval", *Journal of Information Processing and Management*, Pergamon, Vol. 35(4), pp. 559-586, 1999.

[20] KODENSHA, *J-Pekin 2000: Japanese-Chinese*, Chinese-Japanese Twin Translation Software, Kodensha, Japan, 2000.

[21] K.L. Kwok, "Comparing Representation in Chinese Information Retrieval", *In Proceedings of the ACM SIGIR-97*, pp. 34-41, 1997.

[22] J.H. Lee, H.Y. Cho and H.R. Park, *"n-Gram-based Indexing for Korean Text Retrieval", Journal of Information Processing and Management*, Pergamon, Vol. 35(4), pp. 427-441, 1999.

[23] T.A. Letsche and M.W. Berry, "Large Scale Information Retrieval with Latent Semantic Indexing", *Information Sciences – Applications*, Vol. 100, pp. 105-137, 1997.

[24] K. Lunde, *CJKV Information Processing: Chinese, Japanese and Korean Computing*, O'Reilly & Associates, Inc., 1999.

[25] A. Maeda, "Studies on Multilingual Information Processing on the Internet", *PhD Thesis*, NAIST-IS-DT-9761021, Nara Institute of Science and Technology (NAIST), Japan, 2000.

[26] Y. Matsumoto, H. Kitauchi and T. Yamashita, "User's Manual of Japanese Morphological Analyzer, ChaSen version 1.0", *Technical Report IS-TR97007*, Nara Institute of Science and Technology (NAIST), Japan, 1997, (in Japanese).

[27] M.F. Mudawwar, "Multicode: A Truly Multilingual Approach to Text Encoding", *IEEE Computer*, Vol. 30(4), pp. 37-43, 1997.

[28] NAMZU, Namazu, *A Full Text Search Engine*, http://www.namazu.org/

[29] J.Y. Nie, M. Brisebois and X. Ren, "On Chinese Text Retrieval", *In Proceedings of the ACM SIGIR-96*, pp. 225-233, 1996.

[30] J.Y. Nie, J.P. Chevallet and M.F. Bruandet, "Between terms and Words for European Language IR and Between Words and Bigrams for Chinese IR", *In Proceedings of the Sixth Text REtrieval Conference* (TREC-6), pp. 697-710, 1998.

[31] J.Y. Nie and F. Ren, "Chinese Information Retrieval: using character or words?", *Journal of Information Processing and Management, Pergamon*, Vol. 35(4), pp. 443-462, 1999.

[32] J.Y. Nie, J. Gao, J. Zhang and M. Zhou, "On the Use of Words and N-grams for Chinese Information Retrieval", *In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages*, IRAL-2000, pp. 141-148, 2000.

[33] D.W. Oard and B.J. Dorr, "A Survey of Multilingual Text Retrieval", University of Maryland, *Technical Report*, UMIACS-TR-96-19, CS-TR-3615, 1996.

[34] Y. Ogawa and T. Matsuda, "Overlapping Statistical Word Indexing: A New Indexing Method for Japanese Text", *In Proceedings of the ACM SIGIR-97*, pp. 226-234, 1997.

[35] PERGAMON, "Special issue on Information Retrieval with Asian languages", *Journal of Information Processing and Management*, Vol 35. No.4. Pergamon, London, 1999.

[36] B. Rehder, M.L. Littman, S. Dumais and T.K. Landauer, "Automatic 3-Language Cross-Language Information Retrieval with Latent Semantic Indexing", *In Proceedings of Text REtrieval Conference* (TREC-6), pp. 233-240, 1998.

[37] G. Salton and M.J. McGill, *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[38] SANSEIDO, Sanseido's *The Unicode Kanji Information Dictionary*, Sanseido, Japan, 2000.

[39] M. Shibatani, *The Languages of Japan,* Cambridge Languages Surveys, Cambridge University Press, 1990.

[40] R. Sproat, C. Shih, W. Gale and N. Chang, "A Statistic Finite State Word-Segmentation Algorithm for Chinese", *Computational Linguistics*, Vol. 22 No. 2, pp. 377-404, 1996.

[41] C.L. Tan and M. Nagao, "Automatic Alignment of Japanese-Chinese Bilingual Texts", *In IEICE Transactions of Information and Systems*, Japan. Vol. E78-D. No. 1, pp. 68-76, 1995.

[42] TREC-6, *Proceedings of Text REtrieval Conference (TREC-6)*. National Institute of Science and Technology (NIST), 1998. http://trec.nist.gov/pubs/trec6/t6_proceedings.html

[43] UNICODE, *The Unicode Standard, Version 3.0,* Addison Wesley, Reading, MA, 2000. http://www.unicode.org/

[44] E.M. Voorhees, "Variations in Relevance Judgements and the Measurement of Retrieval Effectiveness," *In Proceedings of the ACM SIGIR-98*, pp. 315-323, 1998.

[45] I.H. Witten, A. Moffat and T.C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, Second Edition, Morgan Kaufmann Publishers, 1999.

Computational Linguistics and Chinese Language Processing
Vol. 5, No. 2, August 2000, pp. 87-100

Compiling Taiwanese Learner Corpus of English

Rebecca Hsue-Hueh Shih *

Abstract

This paper presents the mechanisms of and criteria for compiling a new learner corpus of English, the quantitative characteristics of the corpus and a practical example of its pedagogical application. The Taiwanese Learner Corpus of English (TLCE), probably the largest annotated learner corpus of English in Taiwan so far, contains 2105 pieces of English writing (around 730,000 words) from Taiwanese college students majoring in English. It is a useful resource for scholars in Second Language Acquisition (SLA) and English Language Teaching (ELT) areas who wish to find out how people in Taiwan learn English and how to help them learn better. The quantitative information shown in the work reflects the characteristics of learner English in terms of part-of-speech distribution, lexical density, and trigram distribution. The usefulness of the corpus is demonstrated by a means of corpus-based investigation of learners' lack of adverbial collocation knowledge.

Keywords: learner corpus, Taiwanese Learner Corpus of English (TLCE), Second Language Acquisition (SLA), English as Foreign Language (EFL), quantitative analysis, lexical density, collocation.

1. Introduction

A computer corpus is a body of computerized written text or transcribed speech. Computer corpora are useful for a wide variety of research purposes, in fields such as lexicography, natural language processing, and all varieties of linguistics. The first computer corpus made its appearance in the early 1960s when two scholars at Brown University compiled a one-million-word corpus, known as *the Brown Corpus* [Francis & Kucera, 1964]. It contains a wide range of American English texts with grammatical annotation. For decades, this pioneering work was an important source for linguistic scholars who wished to perform quantitative as well as qualitative that is crucial for a broad coverage system. Third, a static WSD model is unlikely to be robust and portable, since it is very difficult to build a single model relevant to a wide

* Department of Foreign Languages and Literature, National Sun Yat-sen University, Kaohsiung, Taiwan
E-mail: hsuehueh@mail.nsysu.edu.tw

analysis of language structure and use [Francis & Kucera, 1982]. In the early 1970s, an equivalent British collection, *the Lancaster-Oslo-Bergn* (LOB) Corpus, was designed and compiled to facilitate comparative studies. Quantitative information on the distribution of various linguistic features in these two corpora became available [Johansson & Norheim, 1988; Nakamura, 1993]. The two corpora and other subsequently compiled corpora are similar in structure and size, and are considered to be first generation corpora.

With the fast development in technology needed for text capture, storage and analysis, the scale of computer corpora has increased considerably, and a corpus of one million words seems to be inadequate for large scale studies on lexis. In the early 1980s, the publisher Collins and Birmingham University compiled the first mega-size corpus, the Cobuild Corpus, for the production of a new English dictionary. The scale of the corpus reached 13-million words by the time the dictionary was published in 1987 [Collins, 1987]. In preparation for a new generation of language reference publications, the corpus was transformed into *the Bank of English* in 1991 and has been growing larger in size ever since. Another well-known mega-corpus, *the British National Corpus*, was compiled between 1991 and 1994 by a consortium of academics and publishing houses. This corpus consists of 100 million words of part-of-speech tagged contemporary written and spoken British English. Access to the corpus was originally restricted within Europe, and it was not until very recently that the corpus was made accessible worldwide. Due to the need for comparative studies of different English varieties as in the first generation, *the International Corpus of English* compilation project was launched in the 1990s [Greenbaum, 1996] to gather written and spoken forms of national varieties of English throughout the world. The project aims to collect up to 20 subcorpora, each containing one million words of English used in countries where English is the first language, and in countries such as India and Singapore where English is an additional office language. The corpus will enable researchers to use each national subcorpus independently for descriptive research and also to undertake comparative studies.

For nearly fifty years, machine-readable language corpora have greatly benefited people in both linguistics and publishing houses. Linguistic scholars have been able to better understand language structure and use with the aid of quantitative data. Publishers have produced new pedagogical tools that reflect the real use of language. However, it was not until the 1990s that scholars in the EFL and SLA sectors began to recognize the theoretical as well as practical potential of corpora and to believe that with the aid of quantitative information, computer learner corpora can form an authoritative basis for obtaining further insights into the interlanguage systems of language learners. Publishing houses also realize the vital role that learner corpora play on designing EFL tools, which can be improved "with the NS (native speaker) data giving information about what is typical in English, and the NNS (non-native speaker) data highlighting what is difficult for learners in general and for specific groups of learners" [Granger, 1998a] However, it is difficult to create learner corpora on the huge scale of native corpora mainly because each collection is usually confined to classroom language.

In 1993 *the International Corpus of Learner English* (ICLE) was launched [Granger, 1993] through academic collaboration worldwide. At present, the corpus contains 14 different national varieties, some of which are subdivided regionally, and each subcorpus contains 200,000 words. A great deal of comparative research has been done based on the ICLE, providing statistics-based interpretation of the learners' lexicon, grammar, and discourse [Granger, 1998b]. Another learner corpus, and probably the largest corpus of single group learners so far, is *the Hong Kong University of Science and Technology Learner Corpus* [Milton & Tong, 1991], which consists of five million words of written English from Cantonese learners. This corpus is intended to be used for the development of English teaching materials in Hong Kong. SLA scholars in Japan soon followed the trend, and several learner corpus projects were launched, such as *the JEFLL corpus* of around 200,000 words from Japanese EFL learners' written data, *the SST Corpus* of 1 million spoken words of learners, and *the CEJL Corpus* of junior high school to university students. In China, *Chinese Middle School Students' Written English* and *Chinese Middle School Students' Spoken English* are two learner corpora forming *the Corpus of Middle School English Education* that was compiled at South China Normal University beginning in 1998. Apart from academic circles, publishing houses such as Longman and Cambridge University Press have also compiled their own learner corpora for the development of their own language related publications.

While many countries around the world have been creating their own learner corpora, little work has been done in Taiwan. *The Soochow Colber Student Corpus* [Bernath, 1998], which was compiled between 1984 and 1995 at Soochow University, can be viewed as a pioneering Taiwanese corpus of learner English. It contains around 227,000 words of written text from junior and senior students of Soochow University and National Taiwan University. No other corpus of comparable size was compiled until 1999 when a one-million-word learner corpus project, *the Taiwanese Learner Corpus of English*, was launched at Sun Yat-sen University. This corpus is a collection of written data from college students majoring in English at the university. The data has been annotated for various linguistic features using the TOSCA-ICLE tagger/lemmatizer [Aarts, Barkema, & Oostdijk, 1997], assigning to each word its lemma and a tag of its morphological, syntactic and semantic information. With the permission of the compiler of the Soochow Colber Student Corpus to incorporate 85% of its contents, consisting of written data from students majoring in English, the scale of the TLCE has increased from its original 530,000 to 730,000 words. The corpus continues to grow in size. Currently, the TLCE is probably by far the largest annotated learner corpus of English in Taiwan. In the following sections, a complete description of the TLCE will be given, including its purpose, design criteria, method of data capture and documentation, corpus structure and grammatical annotations. The quantitative characteristics of the TLCE as well as its pedagogical application will be depicted and illustrated at the end of the paper.

2. Compilation of the TLCE

2.1 Purpose

The history of the computer learner corpus is less than a decade old, but it has been widely considered as "a useful resource for anyone wanting to find out how people learn languages and how they can be helped to learn them better" [Leech, 1998]. Learner output is indeed hard data that SLA scholars can utilize to depict learners' interlanguage systems. The TLCE has been compiled in the hope that it will become a useful resource for SLA scholars who want to understand the internal learning process of Taiwanese learners of English, and in the hope that with corpus-based research findings, EFL teachers will be able to tailor their teaching to students' needs.

2.2 Corpus Design Criteria

It is important to have clear design criteria when compiling a learner corpus because of the heterogeneous nature of learners and learning situations. Clear criteria help make it possible to interpret research results correctly and help justify the results of comparative studies on different corpora.

Table 1 shows the design criteria of the TLCE. The subjects who have contributed data to the corpus are students majoring in English at the three universities, ranging from freshmen to seniors (aged 19 to 22). Their English proficiency varies from intermediate to advanced levels. The TLCE includes written production of two different genres, namely, informal writings and essay writings. Informal writings consist of daily or weekly journals, which the learners are encouraged to keep during their writing courses, and essay writings are the compositions they are asked to submit regularly for their courses. The types of compositions are mainly descriptive, narrative, expository and argumentative.

Table 1. *TLCE Design Criteria*

attributes	
age	19-22
level	Intermediate to advanced
Mother tongue	Chinese
Learning context	EFL
medium	Written text
genre	journals and compositions

2.3 Data capture and documentation

The data of the TLCE are in three forms: electronic files, printouts and handwritten texts. More than half of the collection has been submitted through e-mail, which is the easiest way of gathering data for the corpus. E-mail or Microsoft Word files are converted into text files. Another source of data, learners' printouts, have been scanned and transformed into a machine readable format. Post-editing of the scanned data is

necessary to remove scanning errors. The most time-consuming task is the collection of handwritten texts; all the data have to be keyboarded. As the issue of spelling errors is not a concern in the project and errors would hinder part-of-speech tagging in the subsequent annotation work, all the data in the corpus are spellchecked.

The documentation of each piece of writing is needed for researchers to create their own subcorpora according to selection based on pre-defined attributes, and to carry out different comparison studies. For this reason, details about attributes are recorded as an SGML file header for each text. The information includes the university where the learner is studying, the academic year in which the text is collected, the school year (proficiency level) of the learner, and the genre of the text. For instance, the header

<#nsysu-891-f-DES>

indicates that the text is a descriptive composition written by a freshman at Sun Yat-sen University in the first semester of the 1989 academic year.

2.4 Corpus structure

As stated in Section 2.2, journals and compositions are the two genres of writing collected in the corpus. Journals are informal writings from students, recording what concerns them the most during a day or a week. The journals are sent to their teachers through e-mail systems. Compositions are the essay writings based mainly on different writing strategies: description, narration, exposition and argumentation. The first two are often taught in the first year at universities, whereas the expository and argumentative types are practiced in the second and the third years. Table 2 illustrates the structure of the corpus, including the total numbers of texts and words, and the percentage of the corpus each genre represents.

Table 2. *The Structure of the TLCE*

Text Types	**Total number of texts**	**Total number of words**	**Proportion (%)**
journal	823	213091	29.4
composition			
Description/narration (first year)	435	134363	18.5
		333734	46.1
Exposition/argumentation (second/third years)	738		
		43156	6.0
others	109		

As indicated in the table, the ratio of journals to compositions in the corpus stands at around 3 to 7. Expository and argumentative types of writings are most numerous, making up more than 46% of the whole corpus. Data classified as *others* came originally from *the Soochow Colber Student Corpus* with type labels that did not fit into the TLCE categories. For instance, they are labeled as autobiographical writings, letters, imaginative writings or creative writings.

2.5 Grammatical Annotation

Computer corpora are either raw corpora or annotated corpora. Raw corpora simply contain plain text, whereas annotated corpora have extra encoded features obtained through part-of-speech tagging or syntactic parsing. Part-of-speech tagging is a process of attaching a category and probably other attributes to each word, whereas syntactic parsing provides the structural analysis of each sentence. The former is usually done automatically by rule-based, probabilistic or mixed taggers, and the average tagging accuracy is about 95%; the latter can be done by automatic full/partial parsers outputting one or more syntactic structures for a sentence.

The text in the TLCE is currently part-of-speech tagged using the TOSCA-ICLE tagger [Aarts et al., 1997]. TOSCA-ICLE is a stochastic tagger, supplemented with a rule-based component, which tries to correct observed systematic errors of the statistical components. Each word is given its lemma, and a part-of-speech tag, which consists of a major wordclass label, followed by attributes for subclasses and for its morphological information. There are 17 major word classes in the tag set (see Appendix A) and a total of 270 different attribute combinations.

3. Quantitative Analysis

A major advantage of the corpus approach lies in the usefulness for conducting quantitative analysis. The quantitative features of a corpus provide a basic but global view of the characteristics of the learners' writings. The following findings depict the characteristics of the TLCE as a learner corpus.

3.1 Part-of-speech Distribution

Figure 1 shows the part-of-speech distribution of the corpus. The graph only indicates those parts of speech individually making up at least 5% of the total corpus. As can be seen, Nouns (N) and verbs (VB) exist in similar proportions in the corpus. Pronouns (PRON) are third, followed by prepositions (PREP), adverbs (ADV), adjectives (ADJ), articles (ART) and conjunctions (CONJUNC). Note that the words in nominal form (N or PRON) make up nearly one third of the whole corpus.

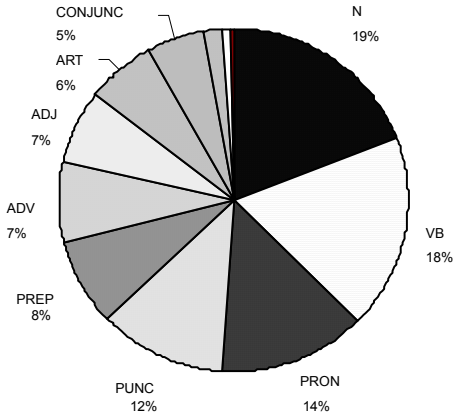


Figure 1. *POS Distribution*

3.2 Type/Token Ratio (Lexical Density)

For open classes, N, VB, ADV, and ADJ, it is desirable to know their type/token ratios. The type-token ratio, also called the lexical density, is often used as a measure of the lexical complexity of a text. Here, it is used as the measure of the word versatility of an open class. It is the ratio of different words to the total number of words in the class and is calculated by the formula

$$Lexical\_Density = \frac{number\_of\_separate\_words(type)}{total\_number\_of\_words(token)} * 100 .$$

Although N and VB have similar distributions as shown in Figure 1,their lexical densities show great discrepancy. As can be seen in Figure 2, the lexical density of N is four times higher than that of VB. This phenomenon is also found in the pair consisting of ADJ and ADV, where ADJ has a much higher density value than ADV. In other words, although the frequency counts of VB and ADV in the learner corpus are similar to those of N and ADJ, respectively, the variety of actual words used in the categories of VB and ADV is much more limited than in the N and ADJ categories.

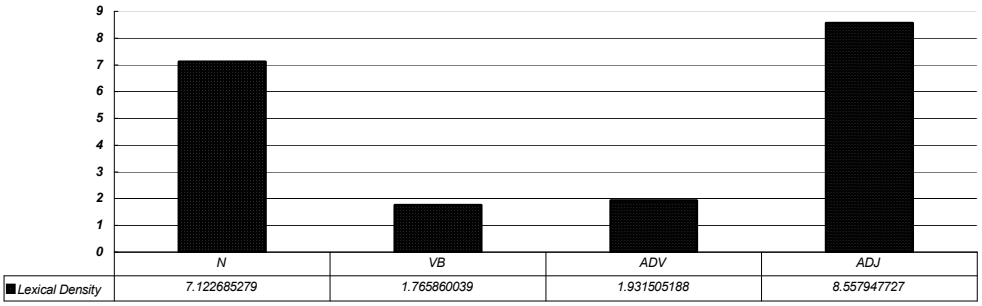


Figure 2. *Lexical Density*

3.3 Part-of-speech trigrams

A POS trigram is a pattern of three adjacent POSs. It reveals to a certain extent the habitual use of syntactic structures by language learners. The corpus has a total of 777,096 trigrams from 2202 different patterns. Hence, the type-token ratio of POS trigrams is as low as 2.8. Table 3 shows the distribution of the front rank trigram patterns according to frequency of use. As can be seen, the first 50 patterns make up a large proportion of use in the distribution diagram. In fact, it is calculated that the top 220 ranking patterns make up 82% of the trigrams. In other words, learners use only 10% of the POS trigram patterns in 80% of their writings. These figures demonstrate the serious lack of structural variations in learners' writings.

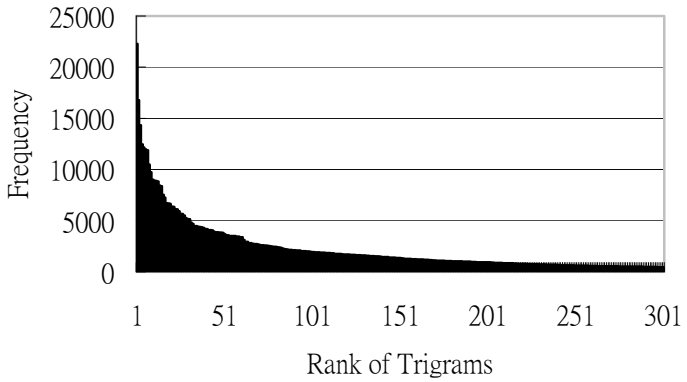


Figure 3. *Trigram Distribution*

4. Pedagogical Application

The main purpose of compiling the TLCE is to provide Taiwanese researchers in the SLA and EFL areas with a large quantity of authentic learner data, which can be used to conduct qualitative analysis based on quantitative information. With the availability of this useful resource, they can utilize advanced corpus analysis tools to systematically uncover the features of non-nativeness existing in learner English. The findings will enable EFL teachers to focus on areas where remedial work is needed. In this section, a pedagogical application of the TLCE is demonstrated through an investigation of learners' lack of adverbial collocation knowledge from both overuse and underuse perspectives. A series of experiments were carried out based on a contrastive approach, comparing learner English (from the TLCE) with native English (from a one-million-word subset of the BNC).

4.1 Top 10 adverbs in the BNC and the TLCE

A frequency list of adverbs with the "-ly" suffix was obtained from each corpus, and their top 10 adverbs were taken into consideration. The left column of Table 3 shows the top 10 adverbs used by the learners, and the right column shows those used by the native speakers. The bracketed number following an adverb indicates the adverb's rank in the other corpus.

Table 3. *Top 10 adverbs in the two corpora.*

	Top NTLCE(learner)	BNC (native)
1	really(1)	really(1)
2	finally(12)	probably(23)
3	usually(4)	*particularly(96)*
4	especially(10)	usually(3)
5	suddenly(l8)	actually(8)
6	easily(I3)	early(22)
7	recently(20)	certainly(27)
8	actually(5)	nearly(32)
9	*deeply(60)*	simply(51)
10	quickly(14)	especially(4)

As can be seen in the table, four of the top 10 adverbs in the TLCE appear in the BNC list, namely, *really*, *usuall*y, *especially* and *actually*. The rest fall into BNC's top 20 group except for *deeply,* whose counterpart is ranked 60. This implies that *deeply* is very overused by Taiwanese learners. By contrast, *particularly*, with the third highest rank in the BNC list, is the one least used by the learners. Sections 4.2 and 4.3 provide a closer examination of these two phenomena, respectively, based on the contexts in which they appear.

4.2 Overuse phenomenon

This experiment examined the context in which *deeply* appears in the TLCE. The adverb can be used to intensify adjectives or verbs. According to the estimation of Mutual Information, the top 10 adjectives or verbs that highly collocate with *deeply* those listed in the left column of Table 4.

The middle and right columns show the adverbs (including *deeply*) which are used by the learners and native speakers, respectively, to intensify words. They are listed in the descending order of their joint frequencies with the corresponding words. As can be seen, *deeply* seems to be chosen most often when learners wish to use an adverb to modify these words, whereas in the BNC, the native speakers use other synonyms (words in bold type) more frequently than *deeply* to intensify the same set of words. *Extremely distressed*, *strongly/greatly influenced*, *greatly impresse*d, *strongly/greatly attracted*, *firmly convinced* and *extremely confused* are collocations that do not exist in the TLCE. This finding suggests that instead of the

monotonous use of *deeply*, Taiwanese learners should be made aware of native speakers' strong preference for the above collocations.

Table 4. *Adverb Alternatives*

Intensified words	**Adverbs in TLCE**	**Adverbs (Synonyms) in BNC**
Distressed	Deeply	extremely, deeply, …
Influenced	deeply, directly, rapidly	strongly, greatly, deeply,…
Moved	deeply, really, suddenly, …	deeply,…
impressed	deeply, especially, really	greatly, deeply,…
attracted	deeply, really, fully	strongly, greatly, deeply,…
convinced	deeply, obsessively	firmly, deeply,…
touched	really, deeply	deeply,…
concerned	deeply, obsessively	deeply,…
confused	deeply, really	extremely, deeply,…
interested	really, deeply, keenly	deeply,…

4.3 Underuse phenomenon

To understand the learners' use of *particularly*, its concordancing lists from the corpora were investigated. There are only 4 instances of the adverb in the TLCE, whereas in the BNC, there are 217 examples. Following is the complete TLCE list and a selected sample of the BNC list:

TLCE concordancing list

First of all, the government, <particularly> the Ministry of Administration,
self-defense, the teachers, <particularly> the elementary school teachers,
is still applied universally , <particularly> in cram schools for high schools
take your words seriously, <particularly> in foreign countries. They might

BNC concordancing list (selected)

ncy food aid in 1990. 'We're <particularly> concerned about the situation in
may have been linked with a <particularly> violent six-week strike by rail
n French international thinking <particularly> over France's role as the motor
ront and other radical groups, <particularly> among the rapidly expanding
he past six months, and many, <particularly> the US, are expected to argue
st and provide grants for artists, <particularly> students, in the region. Thr
strial and social development, <particularly> after Renault was nationalised in
hey still have a very useful role, <particularly> when it is the function of t
the landscape study shown here. 1 <particularly> liked the rounds for their v
hot poker. These colours work <particularly> well in late summer and early

As can be seen in the TLCE concordancing list, there are only two different functions of *particularly* in the learners' writings: it is used to modify either a noun phrase or a preposition phrase. However, there are more functions of the adverb in the native speakers' writings. Apart from noun and prepositional phrases, the native speakers also use it to intensify clauses, verb phrases, adjectives and even

adverbs. Table 5 shows the percentage of each of the grammatical functions used in each corpus. The findings are two fold. First, the learners seem to possess limited knowledge of *particularly*'s grammatical behaviours. Only two out of the six functions are actually found in the TLCE. Second, the learners are not clear about the possible uses of *particularly*. Its collocation with adjectives makes up 42% of the BNC examples, the highest among all, but yet it is not used in this way by the learners at all. The above findings suggest that learners should be informed of the grammatical function of the adverb during the learning process.

Table 5. *Distribution of Grammatical Collocations of "particularly"*

Grammatical collocation	BNC(%)	TLCE(%)
ADJECTIVE	42	-
PREPOSITION PHRASE	28	50
NOUN PHRASE	15	50
CLAUSE	7	-
VERB	6	-
ADVERB	2	-

5. Summary and Outlook

This is the first large-scale tagged Taiwanese learner corpus of English. Preliminary results show several interesting characteristics of the learner corpus in terms of its part-of-speech distribution, the lexical density of its main categories, and the distribution of its trigram structures. An example of pedagogical application has been used to illustrate the usefulness of the corpus. These efforts have been made in the hope that scholars in language education and research will benefit from this pioneer learner corpus, which will be made available soon on website with software tailored to facilitate corpus analysis.

Acknowledgements

Financial support from the National Science Council of the Republic of China of this work under contract No. NSC 89-2411-H-110-024 is gratefully acknowledged. I would also like to express my gratitude to Colman Bernath, the compiler of the Soochow Colber Student Corpus, for allowing his corpus to be incorporated into the TLCE. I am also greatly indebted to the following colleagues for helping me collect data: Dr. Shu-ing Shyu, Dr.Ching-yuan Tsai, Dr. Shu-li Chang, Dr. Shu-Fang Lai, Dr.Yuan-jung Cheng, Hue-jen Wen, Alex K.T. Chung, Chu-jen Loh, Dr. Ting-yao Luo at Sun Yat-sen University and Dr. Zhao-ming Gao at National Taiwan University.

References

Aarts, J., Barkema, H., & Oostdijk, N. 1997. *The TOSCA-ICLE Tagset* . Nijmegen: University of Nijmegen, The Netherlands.

Bernath, C. 1998. *Soochow Colber Student Corpus*. Available: ftp://ftp.scu.edu/tw/scu/english/colber.

Collins. 1987. *COBUILD English Language Dictionary*. London and Glasgow: Collins.

Francis, W., & Kucera, H. 1982. *Frequency Analysis of English Usage:Lexicon and Grammar*. Boston: Houghton Mifflin.

Francis, W. N., & Kucera, H. 1964. *Manual of Information to accompany ' a Standard Sample of Resent-Day Edited American English, for Use with Digital Computers'* Department of Linguistics, Brown University.

Granger, S. 1993. The International Corpus of Learner English. In J. Aarts, P. d. Haan, & N. Oostdijk (Eds.), *English language Corpora: Design, Analysis and Exploitation.* pp. 57-69 Armsterdam: Rodopi.

Granger, S. 1998a. The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on Computer*. pp. 3-18 London and New York: Longman.

Granger, S. (Ed.). 1998b. *Learner English on computer*. London and New York: Longman.

Greenbaum, S. (Ed.). 1996. *Comparing English Worldwise: The Interational Corpus of English*. Oxford: Clarendon Press.

Johansson, S., & Norheim, E. H. 1988. The subjunctive in British and American English. *ICAME Journal* (12), 27-36.

Leech, G. 1998. Preface. In S. Granger (Ed.), *Learner English on Computer* . New York: Longman.

Milton, J., & Tong, K. (Eds.). 1991. *Text Analysis in Computer Assisted Language Learning*. Hong Kong: Hong Kong University of Science and Technology.

Nakamura, J. 1993. Quantitative comparison of modals in the Brown and LOB corpora. *ICAME* (17), 29-48.

Appendix A: part of speech set of TOSCA Tagger

Label	Major word class
ADJ	Adjective
ADV	Adverb
ART	Article
CONJUNC	Conjection
EXTHERE	Existential there
GENM	Genitive marker
HEUR	(unknown)
misc	Miscellaneous
N	Noun
NADJ	Nominal adjective
NUM	Numeral
PREP	Preposition
PROFM	Proforin
PRON	Pronoun
PRTCL	Particle
PUNC	Punctuation
VB	Verb