GlobalMind - Bridging the Gap Between Different Cultures and Languages with Common-sense Computing

by

Hyemin Chung

B.S., Seoul National University, Korea (2003)

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of

Master of Science in Media Arts and Sciences

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2006

Author_____

Program in Media Arts and Sciences
August 11, 2006

Certified by_____

Walter Bender
Senior Research Scientist
MIT Media Laboratory
Thesis Supervisor

Accepted by_____

Andrew B. Lippman
Chair, Departmental Committee on Graduate Students
Program in Media Arts and Sciences

GlobalMind - Bridging the Gap Between Different Cultures and Languages with Common-sense Computing

by

Hyemin Chung

Submitted to the Program in Media Arts and Sciences,
School of Architecture and Planning,
on August 11, 2006, in partial fulfillment of the
requirements for the degree of
Master of Science in Media Arts and Sciences

Abstract

The need for more effective communication across different countries has increased as the interactions between them have been growing. Communication is often difficult because of both language differences and cultural differences. Although there have been many attempts to meet the communication need on the level of language with machine translators and dictionaries, many problems related to cultural and conceptual differences still remain.

To improve traditional machine translators and cross-cultural communication aids, it is necessary to develop automated mechanisms to analyze cultural differences and similarities. This thesis approaches the problems with automatic computation of cultural differences and similarities.

This thesis, GlobalMind, provides common-sense databases of various countries and languages and two inference modules to analyze and compute the cultural differences and similarities from the databases. I describe the design of GlobalMind databases, the implementation of its inference modules, the results of an evaluation of GlobalMind, and available applications.

Thesis Supervisor: Walter Bender
Title: Senior Research Scientist, MIT Media Laboratory

GlobalMind - Bridging the Gap Between Different Cultures and Languages with Common-sense Computing

by

Hyemin Chung

The following people served as readers for this thesis:

Thesis Reader_____

Henry Lieberman
Research Scientist of Media Arts and Sciences
MIT Media Laboratory

Thesis Reader_____

Sung Hyon Myaeng
Professor of Computer Science
Information and Communications University, Korea

Acknowledgements

First I would like to really thank my parents, Kyu Soo Jhung and Kwang Hwa Chung, who always have and continue to support me, believe me, guide me, and love me.

I would like to thank my advisor Walter Bender, for his guidance and support, without which this research would not have been possible. I also thank my thesis readers Henry Lieberman and Sung Hyon Myaeng, for their valuable feedbacks and comments.
Additionally, I'd like to thank the late Push Singh. His work and the discussions with him inspired this research. Thank you, Wonsik Kim, who helped me a lot in implementing this project. Thanks to Masanori Hattori and Edward Shen for their contributions to GlobalMind. Thanks to all the GlobalMind participants.

Thanks to all the members of Electronic Publishing group and Common Sense Computing group, Ian Eslick, Benjamin Mako Hill, Bo Morgan, Dustin Smith, and Scott Vercoe. Together, they filled my two years at Media Lab with enthusiastic discussions and inspiration of exciting ideas. Thanks to many people in Media Lab for the passionate atmosphere we made together, especially Jackie Chia-hsun Lee, my friend and collaborator.

Friends at MIT, S&P girls, and friends in Korea, thank you for painting my two years at MIT with colors of delight. Thank you Kyungbum Ryu for everything. You're the best.

I thank Samsung Lee Kun Hee Scholarship Foundation for supporting me during my master's program.

엄마 아빠 감사합니다. 그리고 사랑해요.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Difficulties of Cross-cultural Communication

In these days, the number and the scale of multinational organizations have been increasing, and the interactions among countries have become more frequent. Although these changes have increased the need of effective cross-cultural communication, it remains difficult because of both cultural and language differences.

In this section, I will describe the challenges of cross-cultural communication and the limitations of traditional automated mechanisms for cross-cultural communication, mainly the machine translators.

1.1.1 Cultural Differences

In cross-cultural interactions, people should consider and understand the cultural background of each other in order to have successful interactions [1]. Much research has shown that in cross-cultural communication and negotiation the cultural differences between communicators affect the outcome of the negotiation. In the negotiation among different countries, small misunderstandings caused from cultural differences lead the whole negotiation

to bad results [37]. Herring [18] showed cross-cultural counselors should understand cultural differences and apply those differences to their non-verbal communication styles to avoid misunderstandings. Condon [7] emphasized the importance of understanding cultural differences relative to that of understanding language differences in that the misunderstandings from language differences could easily be recognized but misunderstandings from cultural differences could not easily be deciphered and corrected. Thus, the consideration of cultural contexts in cross-cultural communication is essential to successful interactions. However, a systematical method to automate analysis of cultural differences has not been completed yet.

In this section, some of cultural differences in cross-cultural interactions will be discussed.

Expected Behaviors

In human interactions, it is important to know what is expected of each other. People are expected to behave in certain ways in certain situations, and the failure of behaving in the expected manner is often regarded rude and impolite.

For example, when a person goes to an American shop, he expects someone to help his shopping, helping him find a product in the shop and suggesting which products would be the best for his purpose. If the attendant gives a map of the warehouse to the customer and asks him to search for it by himself, the customer will be infuriated.

The expected behaviors differ by culture; people in different cultures will expect different things even in the same situation. Thus, it is important to understand the different expectations of the different cultures to avoid misunderstandings and misled communication.

Contexts of Communication

Communication is exchange of expressions and signals. These exchanged expressions and signals are received by a person and analyzed with background contextual knowledge the listener has. Thus, differences in the contexts cause the differences of perception of the expressions. If the listener has different contextual knowledge, he can misunderstand the point that the speaker was intending to make, making communication difficult.

differences in the contexts cause the differences of perception of the expressions. Without considering the differences of the contexts, the mis-perception can cause misunderstandings, and lead the communication to unintended result.

Two persons are talking about stock markets. One person from America says "I saw a lot of red arrows." He means stock prices went down because in American culture "red arrows" in stock markets are used to mark stocks that have gone down. However, the other Korean person will understand the remarks as the stock prices went up because in Korea, people use red arrows to express rising stock prices.

The contexts people have are different by many factors. However, the cultural background of the people is one of the biggest factors for the context knowledge they have. Thus, for successful cross-cultural communication, the differences of context knowledge should be considered.

1.1.2 Language Differences

How many languages are in the world? While nobody can really answer the question accurately, Ethnologue Languages of the World [12] lists 7,299 primary names of languages. Considering that this list does not count the similar languages used in different regions such as American English, British English, and Australian English, the number of different languages is larger than that.

The language difference is one of the most well perceived and thoroughly researched problems in cross-cultural communication. Because it is almost impossible to communicate with other cultures without solving the language difference problems, the language differences have been researched and studied by many people from linguistic researchers to elementary school students. Such wide-spread research has resulted in automated mechanisms such as machine translators.

The efforts to solve the language difference problem with automated mechanisms have resulted in many different kinds of mechanisms of machine translation. Knowledge-based machine-translation systems use parallel bilingual-knowledge databases [23] [42] and statistical machine-translation systems use bilingual corpora [4] [44]. Lastly, example-based machine-translation systems use parallel bilingual-example databases [5] [41]. While these mechanisms have solved many parts of the problem, there still remain problems, and many of these remaining problems, which will be described in this section, cannot be solved without consideration of cultural differences.

There have been many discussions about if an accurate translation between two different cultures is possible [38]. It remains difficult to make an accurate translation between two cultures; in many cases, a vocabulary or an idiom in one culture is not found in another culture; and even if a similar vocabulary exists, it does not mean the same experiences when the cultural backgrounds are different [39]. Thus, it is necessary to consider the cultural differences as well as the language differences when translating languages.

In this section, I describe the problems of language differences which can be improved with understandings of cultural differences.

Untranslatable Concepts

Munter [32] pointed that English does not have a word for Korean word “KI BUN” which has similar but different meanings to “inner feelings of one person” or “mood.” Similarly,

English word “mind” is not well translated into French. Thus, it is difficult to translate Korean or English speeches with the word “KI BUN” or “mind” to English or French, respectively.

Although this problem is grounded in language differences, it cannot be solved without understanding each other’s cultures. The existence or absence of the word is also closely related to the existence or absence of the concept itself. There is the high likelihood that Americans do not have the concept of “KI BUN” or, if they have it, they do not consider “KI BUN” as important, while Korean people always care the “KI BUN” in the interactions with other people. Thus, to understand the word “KI BUN,” it is necessary to understand the concept and the cultural background of the word.

Indistinct Conceptual Borders

Most of the current machine-translation systems and supporting corpora tools index the meanings of words to solve ambiguity problems. This approach can work well when it is limited to specific domains, in which the border of concepts is clear and users in different languages are sharing, and should share, exactly the same concepts, such as in the case of technical documents. However, for traditional concepts found in the activities of our daily lives, concepts are often not clearly distinguished from one another, so it is challenging to delineate between concepts. For example, the English noun “plant” is categorized in the Merriam-Webster OnLine dictionary into four meanings, in Dictionary.com into five meanings, and in the Oxford English dictionary into eleven meanings [8] [29] [35].

For a similar reason, using word-to-word mapping is problematic. In most machine-translation systems, indexed words are mapped to indexed words between target languages. In addition to the indistinct conceptual-border problem described above, words in different languages often have similar but different conceptual borders. In American English, a spoon is an oval concave utensil made of metal for soup or tea, which are not main dishes. In Korea, a “spoon” is a round and much flatter utensil made of metal for a main dish. A Japanese

"spoon" is made of ceramics and is rarely used. In this case, although those three utensils have similar concepts, being used for liquid food, their concepts are different in shape, material, and uses.

(a) Korean "spoon" (b) Japanese "spoon" (c) American "spoon"

Figure 1-1: "spoons" in different cultures

Different Uses of Expressions

All expressions have their uses, but the uses of expressions are different from culture to culture. Thus, some expressions with different meanings can be used for the same uses in different cultures, and the other expressions with the same meanings can be used totally differently between cultures.

For example, Americans often say "sure" in response to "thank you" or "I'm sorry" while Korean people often say "A NI E YO(no)" in response to thanks or apologies. "Sure" and "no" have almost opposite meanings, but in this situation, they are used for the same uses. Without understandings of these different uses of expressions, when Korean person visits America and hears "sure" in answer to "thank you" or "I'm sorry" he feels as if he is derided because he misunderstands the expression as "yes, you should surely feel sorry for me" or "yes, you should surely thank me."

1.2 GlobalMind Design Goals

As discussed above, cross-cultural communication needs much consideration of cultural backgrounds. Although people have recognized the importance of consideration of cultures, it has been difficult to use the cultural contexts in machine translators or other automated cross-cultural communication tools.

The work in this thesis is designed to provide programming tools for analyzing cultural contexts to reduce the problems described above and to improve the quality of cross-cultural communication. GlobalMind provides the large-scale databases of several different cultures and languages and the analysis modules of the databases. GlobalMind is designed to support other communication-aid tools such as machine translators.This section will describe the design goals and the major features of GlobalMind.

1.2.1 Automated Mechanisms for Cultural Contexts Analysis

As described in section 1.1.1 the cultural differences should be considered in the cross-cultural communication assistant tools. For this task, it is essential to have an automated mechanism to analyze cultures and to extract the similarities and the differences between cultures. [2] showed a possibility of assistant programs to improve the understanding of cultural differences. However, it has limitation in both depth and breadth of data the and uses because the database of the differences between two cultures are entered manually by human, not automatically computed. With manual input, it is difficult to extend and generalize the work.

In this thesis, GlobalMind provides two inference modules: Similar-concept Inference Module and Differences Inference Module. These inference modules extract the similarities and the differences between two cultures automatically. With this automated mechanism, the comparison and analysis of cultural differences can be used by any other programs and can be easily extended to other languages and to various kinds of applications.

1.2.2 Easily Enlarged and Resilient Multilingual/Multicultural-Text Database System

To analyze cultures and languages, it is necessary for GlobalMind to know about the cultures and the languages, which means to have data about them. Thus, one of the goals of GlobalMind should be building an easily enlarged and resilient database system with the knowledge of different languages and different cultures.

Because GlobalMind culture/language analysis modules works on the database, the quantity and the quality of the database is critical for the best result. However, it is hard for just a few people to build a database with enough entries and detailed context, continually updated to accommodate a changing world. Therefore, I re-used the Openmind data-acquisition system for GlobalMind data acquisition. The Openmind common-sense database gathered common-sense knowledge from Internet volunteers; it gathered more than 400,000 common-sense assertions from 1999 to 2002 [40], and more than 700,000 items as of November 2005 [34]. The database has detailed contexts for each item; all the items are related to each other, and related items form the contexts of each item. The knowledge in the database is expanded by Internet volunteers, so it can reflect changes in the world.

1.2.3 Context-Based Analysis

As described in Section 1.1.1, understanding contexts is an important key for successful cross-cultural communication. To improve context analysis, GlobalMind uses context-based approach. Here, I use the term "context" not as limited to the domain of given words or their sentences, but also expanded to all the related associations of the words. For example, the context of the word "shampoo" includes "used while taking a shower," "used on hair," "followed by rinse," "good fragrance," etc.

Not only different associations or related information but also conceptual border differences and cultural differences can be represented with different contexts. For the "spoon"

example above, the word "spoon" in the USA, Korea, and Japan will have different contexts; the "spoon" in the USA will have the context of "soup" and "tea," the "spoon" in Korea will have the context of "metal" and "main dish," while "spoon" in Japan will have the context of "ceramic."

GlobalMind uses a networked database of common-sense taken from various cultures and languages to apply this context-based method, where the context of the language is represented by common-sense knowledge.

1.2.4 Relation-to-Relation Mapping

To fully support the context-based approach, relation-to-relation mapping is required over word-to-word mapping. At first, there are many words which do not have exactly the same matching words in other languages or exactly the same contexts. Word-to-word mapping ignores differences in contexts. Moreover, the mapping among the words will not change even if the contexts of the words change.

Here I used relation-to-relation mapping rather than word-to-word mapping. For example, mapping between an English relationship "tree-KindOf-plant" and a Korean relationship "NA MU(tree)-KindOf-SIK MUL(plant)" is more suitable than a mapping between an English word "plant" and a Korean word "SIK MUL(plant)."

This approach has another potential advantage: word-to-word mapping systems always require a disambiguating process to map the words to each other and to find the mapped words for a given word. When the sentence "a tree is kind of a plant" is given, the word-to-word mapping should disambiguate the word "plant" to know whether it is about living organisms or buildings before searching for the corresponding Korean word "SIK MUL (plant as living organisms)" or "GONG JANG (plant as buildings)." However, relation-to-relation mapping does not need a disambiguating process because intrinsically it is unambiguous mapping.

1.3 Extended Examples

In this section, I will describe the situations where the problems caused from cultural differences between American cultures and Korean cultures impede communication, and show how GlobalMind can improve the situations.

1.3.1 Scenario One

When American people say "there is a party at my home, please bring your own beer," usually it does not mean that you should bring beer and no other beverage. Beer is one of the most common alcoholic beverages in the USA, and the sentence "bring your own beer" means "bring your alcoholic beverage" with beer as a symbol of common alcoholic beverage. However, if the sentence is directly translated from English to Korean without understanding this cultural background, Korean people will misunderstand that they must bring beer even though they don't want to drink beer. Another example is "forks and knives." Forks and knives usually symbolize main utensils for main dishes, which are similar to spoons and chopsticks for Korean people rather than the forks and the knives which are not frequently used in Korea. Thus, when the sentence is translated, there should be consideration of cultural understandings.

GlobalMind application will help this situation with consideration of cultural context of each phrase. In the example of "bring your own beer," GlobalMind can infer that "beer" is a kind of common alcoholic beverage in American party cultures, which is similar to "SO JU(Soju, Korean gin)" in Korea. By this way, GlobalMind cultural dictionary can help users to understand the real meaning of the sentences.

1.3.2 Scenario Two

When a certain situation is given, people are expected to react in certain ways. If people do not behave in the expected ways, on many occasions it embarrasses other people and

is regarded as rude or impolite. However, these expectations are different for each culture. In the same situation, a person's behavior which is perfectly acceptable in one culture can be incivil, rude, or even a sacrilege in another culture. Thus, foreign visitors should be aware of cultural differences before they behave in a way of they do in their own countries. However, it is not easy for visitors to know every detail of cultural differences, and it may lead them to unintentional mistakes. Here let me illustrate an example.

A nice and gentle businessman from Korea visited the USA. He went to a restaurant. Although he saw the hostess at a small front desk by the door, he just ignored her, went in to the restaurant, and grabbed an empty table near the window. A waitress came to him, said hello, gave him a menu, and went back to the kitchen. He read the menu and was ready to order. He looked around and found a waiter. He raised his hand and shouted "here!" The waiter came to him and the businessman ordered his food. After eating all the food, he went to the front desk and asked how much it was. The hostess answered it was $20 including tax. He paid $20 with his credit card and went out to the street.

He made three rude mistakes. At first, he should have waited for the hostess at the front desk to escort him to his table rather than grabbing his table by himself. Secondly, he should have waited for his own waitress to order rather than calling on other waiters. Lastly, he should have paid tips in addition to the cost of the food and tax. These mistakes are caused not because he was a rude person but because he did not know cultural differences about restaurants between Korea and the USA. In Korea, his behaviors are adequate while they are inappropriate in the USA. Thus, these problems can be solved, or at least improved, if there is an assistant who can figure out what the differences between the two cultures in certain situations are.

Here let me hypothetically show how GlobalMind and its application can solve these problems in the above situation.

A nice and gentle businessman from Korea visits the USA. He goes to a restaurant. When

he enters the restaurant, he checks his cellular phone with GlobalMind Intercultural Assistant Program. The GlobalMind program compares Korean cultures and American cultures related with a restaurant and finds that there are three major differences. The program prompts him that in America he should wait at the front desk to be seated. Thus, he stands at the front desk, and a restaurant manager at the front desk guides him to his table. The program also shows him that there is a waiter or waitress who will serve him this night, and it is not a good behavior to call and order to waiters other than his own waiter or waitress. Thus he waits for his waiter to serve him instead of beckoning other waiters. Finally, when he finishes his meal the program shows him that it is common to pay for the meal at his table and that although it is not written in the bill, he should pay not only the price and the tax of the meal but also a tip which is usually 15% of the whole price of the meal. Thanks to the program, he does not make any rude mistake and finishes his dinner with smiles.

Chapter 2

Background and Related Work

The research in this thesis is focused on finding cultural similarities and differences between large common-sense knowledge databases in different languages. To our knowledge, this problem has not been attacked directly by other research.

First of all, the appearance of very large Commonsense knowledge bases is quite recent. The three most developed such resources are Open Mind Common Sense, Cyc [25], and ThoughtTreasure [31]. Cyc has collected knowledge only in English, and with little thought to cultural differences. Cyc does have a mechanism for establishing contexts [25] and context-dependent inference, but it has not been used, so far, for relativizing inference to cultural contexts. ThoughtTreasure has some bilingual knowledge in English and French, but all such knowledge has been hand-crafted by the author. It has no automatic method for establishing new cultural correspondences and cultural analogies.

As already described in Section 1.1.1, there is an extensive literature on differences between cultures and languages. These have been discussed in the context of interface design and of machine translation. We will discuss these below.

Artificial Intelligence has long considered techniques for mapping between different seman-

tic networks expressing different kinds of knowledge. We will discuss related work below in the areas of Ontology Alignment and in Analogy.

2.1 Cultural Issues and Interfaces

Aaron Marcus [28] and others have written extensively on the need for cultural sensitivity in user-interface design. Many people including Russo and Boor [36], and Khaslavky [22] suggested the design strategies with cultural consideration. But they have only implored human user-interface designers to familiarize themselves with cultural differences and take them into account in designing interfaces, particularly to use visual representations that are meaningful to a given culture and audience. They have not worked on directly representing cultural knowledge in the machine and having the machine compute cultural differences automatically, on which GlobalMind focuses.

There is also much work in internationalization of interfaces [43]. This involves translating text used in interfaces into different languages. The bulk of this work is concerned with separating the parts of the interface that are dependent upon language and culture from those that are not. Again, there is usually no provision for explicitly representing cultural assumptions or automatically translating cultural knowledge from one language to another.

2.2 Machine Translation

It has been long known that cultural differences play an important role in machine translation. A general reference on natural language processing that covers machine translation is [21]. Many mis-translations occur from the lack of Commonsense knowledge, or from inappropriately carrying cultural assumptions from one language to another.

The most important problem in language translation affected by cultural assumptions is

Word Sense Disambiguation, discussed below. Single words tend to have several senses, and choosing the correct sense to translate a foreign word requires some consideration of the context of the word. Commonsense knowledge is an important source of context that is not usually explicitly considered in the natural language literature.

In addition to Word Sense Disambiguation, implicit context plays an important role in natural language understanding and translation. The importance of collecting and employing Commonsense knowledge is to make explicit that implicit context. Much interpretation of natural language depends on metaphors [24]. Metaphors can be considered as generalizations of Commonsense situations, we show in this thesis how these generalizations can be carried over from one language and culture to another.

2.3 Word Sense Disambiguation

As explained above, choosing the right word sense during translation poses some difficulties. WordNet [14] is a computational lexicon that carefully distinguishes between word senses. Versions of WordNet also exist in other languages. However, WordNet by itself has no mechanism to choose between the various senses, nor to map word senses in one language to those in another language. Much work has been done on using statistical measures such as lexical affinity and latent semantic analysis [27] as representations of context, to use in word sense disambiguation.

2.4 Ontology Alignment

Ontology Alignment is an active area of research in Artificial Intelligence [13] [33]. The idea of Ontology Alignment is to figure out how to map one conceptual hierarchy onto another, given that the two hierarchies may have been developed independently. Like our inference modules for figuring out similarities and differences between languages and cultures, Ontology Alignment also computes similarities and differences. But OA is limited

to definitional knowledge and formal subsumption hierarchies, rather than our contingent common-sense assertions. Cross-language and cross-cultural OA has also remained difficult.

2.5 Analogies

Finally, much work in Artificial Intelligence concerns analogies. The classic reference is [15]. Gentner's Structure Mapping theory emphasizes, as we do, coordinated mapping of the topology of relations rather than single words or concepts. It has not yet been applied across languages and cultures, nor has it taken advantage of a Commonsense knowledge base rather than small, limited formal logic representations.

Hofstader [20] presents a delightful tour of the importance of analogy and metaphor in language translation. In [19], he and his colleagues explore computational and statistical mechanisms of analogy.

Chapter 3

Design and Theory

The main two components of GlobalMind project are database of common-sense of the world and the inference algorithms and modules to process the accumulated data. In this section, I will illustrate the design of GlobalMind data structures and then describe the algorithms to process the data.

3.1 Design of Data Structure

GlobalMind data are a complicated network of networks of common-sense database of each country. Common-sense knowledge is connected with other common-sense knowledge. Thus, common-sense data of each country form a complicated network. Liu [26] established a form of common-sense network, ConceptNet, and showed decent results with the network form. GlobalMind also uses a similar common-sense network for the network of each country. And then the common-sense knowledge of one country is connected with common-sense knowledge of another country, establishing connections between networks.

In this section, I will describe the structure of the networks from the smallest unit to the largest unit.

3.1.1 Node

A node is the smallest unit in the GlobalMind database. One node represents one concept. One node may consist of one or more words. For example, "student" or "school" as well as "wake up in morning" and "drive fast" can be nodes. A node is combined with another node through a link and become a predicate.

3.1.2 Link

A link is the relationship between two nodes. A link has the direction, which shows the link starts from which node and ends at which node, and the relationship, which shows the kind and strength of the relation between two nodes. The link "–>LocatedAt–>" means the left node is located in the right node and the link "<–IsA<–" means the right node is a kind of left node. GlobalMind adopted 22 different kinds of relationships for links from ConceptNet [6]. The Table 3.1 shows the kinds of relationships GlobalMind is using.

3.1.3 Predicate

A predicate is a combination of two nodes and the link between the two nodes. One predicate contains one common-sense datum, and thus it is the basic unit of GlobalMind database to process and analyze common sense. In this thesis if I refer the size of database or the number of common-sense items, it means the number of predicates.

For example, a node "student" and a node "school" combine with a link "–>LocatedAt–>" and form a predicate "student–>LocatedAt–>school" which means the common-sense that a student is usually found at a school.

3.1.4 Network

A network is a set of predicates in one language and a country (or region). GlobalMind assumes that if two groups have different languages or if they are included in different coun-

Table 3.1: Relationships used in GlobalMind links

relationship	"A–>relationship–>B" means	example of A	example of B
CapableOf	A can do the activity of B	anteater	eat ant
DefinedAs	A is defined as B	prince	son of king
DesireOf	A desires B	people	live
DesirousEffectOf	A makes someone wants B	hunger	eat food
EffectOf	A makes effects like B	stay up late	wake up late
FirstSubeventOf	B happens first while doing A	take shower	turn on water
InstanceOf	A is an instance of B	MIT	university
IsA	A is a kind of B	apple	food
LastSubeventOf	B happens last while doing A	write paper	hand in
LocationOf	A is in B	snow	Boston
LocationOfAction	A is done in B	study	school
MadeOf	A is made of/from B	cup	plastic
MotivationOf	B is the motivation of A	eat	hunger
NotDesireOf	A does not desire B	people	die
OnEvent	On A, B happens	funeral	mourn
PartOf	A is a part of B	wheel	car
PrerequisiteEventOf	B should be done before A	eat	wash hand
PropertyOf	A has characteristics like B	snow	white
SubeventOf	B happens while doing A	eat	chew food
SymbolOf	A is a symbol of B	dove	peace
ThematicKLine	A reminds B	keyboard	mouse
UsedFor	A is used for B	computer	get information

tries, then they are different cultural units. A network represents the common sense of one cultural unit.

Because nodes in one predicate can be shared with other predicates, as the number of predicates is increased more links and connections are established among nodes. Thus, when we gather the predicates of one culture, the predicates form a complicated graph where predicate's nodes are used as nodes in the graph and links are used as edges in the graph. The graph of predicates of each culture is called as a network in GlobalMind. Thus, each culture/language has one network with numerous predicates.

The Figure 3-1 shows a part of a GlobalMind network.

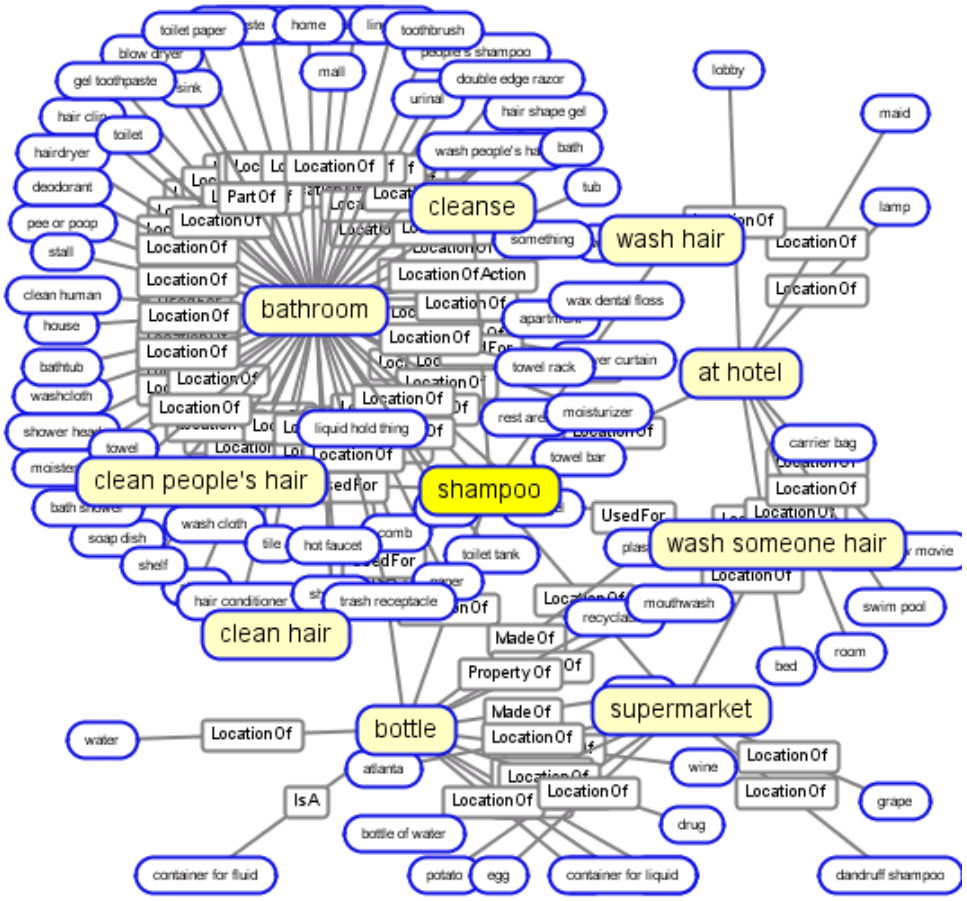


Figure 3-1: Snapshot of GlobalMind network about shampoo

3.1.5 Global Network

Because we are more interested in interactivities between/among cultures rather than activities within one culture, GlobalMind provides a larger network to show the relationships among networks of each country.

One predicate in one language/culture network can be connected with another predicate in another language/culture network. For example, a predicate "tree–>KindOf–>plant" in English/America network can be connected with a Korean predicate "NA MU(tree)–>KindOf–>SIK MUL(plant)." This connection can work as a link between two networks. The networks and these kinds of connections between networks form a larger network. The large network contains the connections between predicates in different countries in addition to all the predicates in GlobalMind. Figure 3-2 shows the concept of a network of networks, the final form of GlobalMind database.

3.2 Inferences

While the GlobalMind database provides the data to be processed, the inference modules are used to process them to make meaningful results. Here GlobalMind presents two different kinds of inference algorithms to find similarities and differences between two cultures/countries.

3.2.1 Similar-concept Inference Module

In cross-cultural communication, it often happens that one person uses a concept but the other person misunderstands it because the concept is used differently in two cultures. To avoid this kind of misunderstandings, GlobalMind provides the inference module to find the most similar concepts between two cultures/languages.

The GlobalMind Similar-concept Inference Module is novel in that it enables a context-based

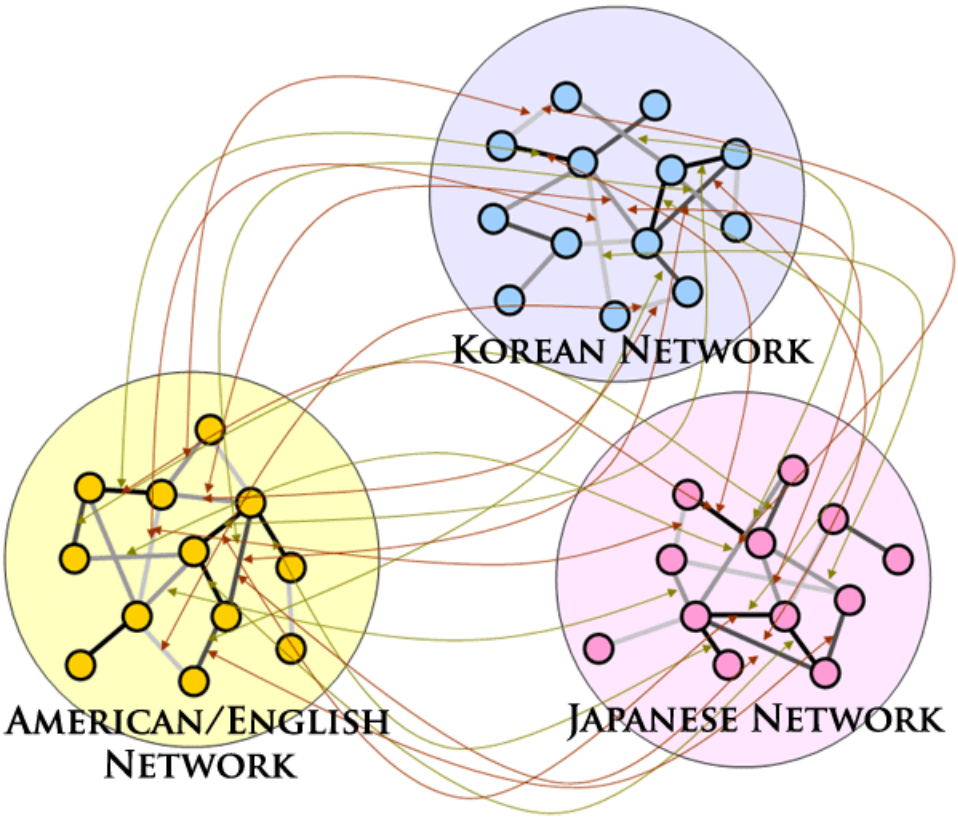


Figure 3-2: Conceptual image of GlobalMind global network

approach rather than a word-meaning-based approach to the problem of word matching.

GlobalMind uses an expand-and-contract method to find the matching link or node for a particular link or node: (1) the context of the given node/link will be browsed by expanding its concept to its neighbor nodes and links and generating a sub network originated from the given node/link with different weight; (2) the context of the given node/link will be found in the target language-based on the existing connections built by bilingual volunteers, it will infer the matching sub-network in the target language and score the correlation of each node and link of a target sub-network; (3) the target sub-network will be contracted into the target node/link based on the scores. Thus, the given node/link and the inferred node/link will have a similar context such as their uses, properties, or locations, even though their meanings in dictionaries could differ. Figure 3-3 shows the concept of the expand-and-contract method.

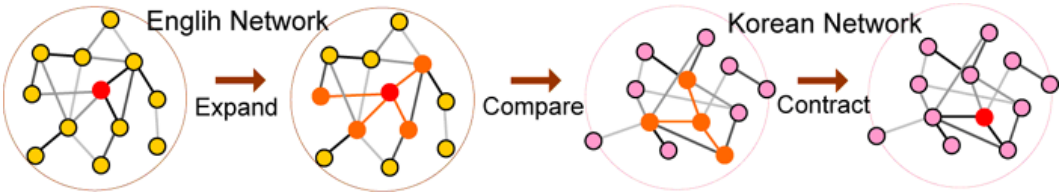


Figure 3-3: Expand and contract method

3.2.2 Differences Inference Module

Understanding cultural differences is important to avoid misunderstanding each other's intention and making rude mistakes. Many books about cross-cultural communication teach their readers to know the cultural differences before their readers go to other countries and make a conversation with people in different countries.

The GlobalMind Differences Inference Module is used for comparing two different cultures and finding the differences between the two cultures when there is a given situation. Although there were several attempts to approach the cultural difference problems [2], GlobalMind is different from them in that GlobalMind automatically extract the differences by

comparing the common-sense databases of each culture while other approaches used manually built databases about cultural differences. Thus, GlobalMind can be easily extended to any pair of two different cultures.

GlobalMind uses a compare-and-remove method to find the differences between two cultures: (1) with a given situation, the related common-senses about the given situation will be browsed in both cultures' networks by extracting the networks around the given situation's node; (2) the extracted sub-networks will be compared with each other sub-network; (3) if there is shared or duplicated common-sense in two sub-networks, the shared common-sense is removed; (4) after comparing and removing, remained sub-networks are cultural differences between two cultures about the given situation.

Chapter 4

Implementation

In this section, I describe the details of implementation of GlobalMind project: database design, data acquisition method, data processing, and inference modules.

4.1 Database Design

As described in Section 3.1, GlobalMind database is a network of networks of common sense from the world. Once gathered, common-sense knowledge is reformed as a part of networks, nodes and relationships between nodes. Each node represents a single concept, represented by a word or a phrase; nodes will be connected with a directed relationship between two nodes. For example, "taking a shower" and "getting clean" are nodes, and "EffectOf" is a relationship between these two nodes. Two nodes and the link between the two nodes form a predicate, which is a small unit of meaningful common sense. A collection of all the predicates of one culture/language is a network. The final form of GlobalMind is a collection of networks of common-sense knowledge.

Because of the size and the complexity of data structure, I needed to find an adequate data storage tool. MySQL Database system is selected for the data storage of GlobalMind

because it is efficient and optimized to handle huge amount of data, it has a lot of supports and application modules written by programmers all over the world, and it is freeware.

The simple diagram of database design is shown in Figure 4-1. Because the whole SQL commands for the creation of database tables are attached as Appendix B, in this section I will only describe the most important three data tables.

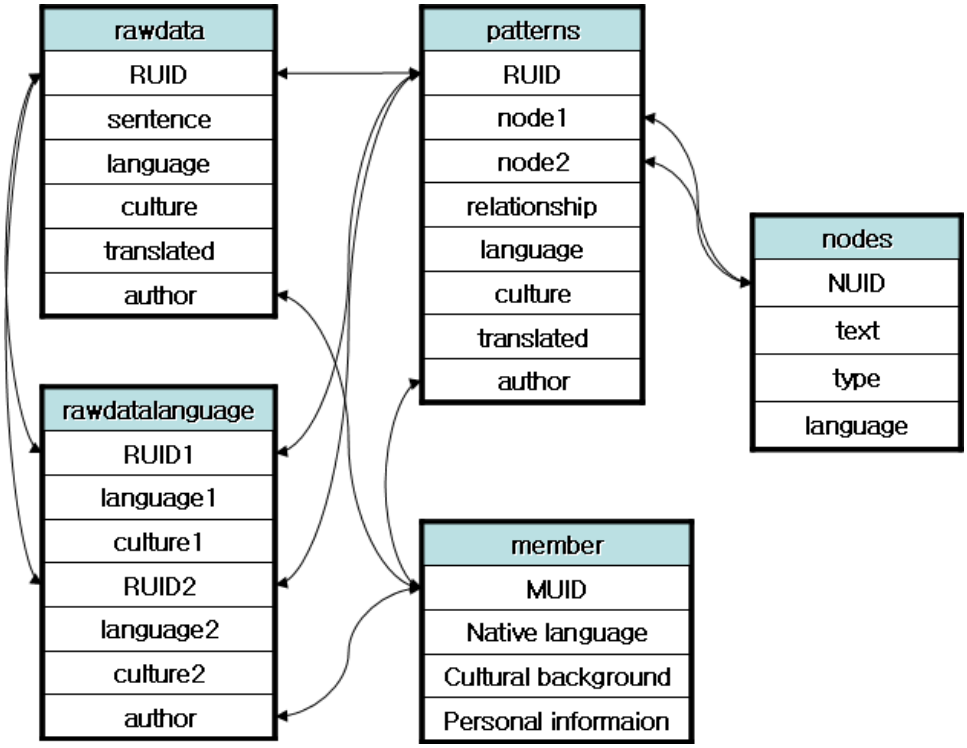


Figure 4-1: Simple diagram of database design

The GlobalMind data is gathered and stored as a form of sentences. For example, "you can find a student in a school" and "people do not want to die" are examples of GlobalMind data forms. The data is processed into a network form, a predicate, like "student–>LocatedAt–>school" where "student" and "school" are nodes and "–>LocatedAt–>" is a link between the nodes. In this process we cannot avoid losing some information. Thus, GlobalMind stores the raw sentence data as well as the processed predicates which are actually used by computer programs and applications.

The original sentences are stored as a form of natural sentences such as "you can find a student in a school" in the "rawdata" table with unique ID, RUID. The table keeps the information of languages and cultural backgrounds of each sentence. Also, if the sentence is not originally written by a user but translated from another language, it also notes the fact that it might not be common sense in this language/culture because it is translated from another sentence in different language/culture.

The processed data like "student–>LocatedAt–>school" is stored as a form of two nodes and one link in the "patterns" table with RUID. The "rawdata" table and the "patterns" table share the same RUID to track each other. The data in this table also contain the languages/cultures information.

Because we are interested in seeing the correlation among different cultures and languages, there should be a database table to store the information about connections among predicates in different languages/cultures. The "rawdatalanguage" table contains bilingual/bicultural connection information. If a predicate with RUID 135 in English/America is connected with a predicate with RUID 245 in Korean/Korea, the table stores a record like "135,eng,usa - 245,kor,kor."

4.2 Website for Data Acquisition

GlobalMind accumulates common-sense knowledge by aggregating the efforts of online websites that are launched in different countries and languages.

There were several attempts to gather large amount of common sense knowledge before GlobalMind project. One of the attempts was OpenMind project. The OpenMind project used a website to gather common-sense knowledge from volunteers of all over the world [40]. OpenMind website was designed to gather large amount of common sense knowledge as a form of sentences. The users of OpenMind could type in their common-sense assertions, and the typed sentences were processed and stored into the internal data storage. To help

users, OpenMind website had several different kinds of activities and templates. For example, users could fill in blanks in templates like "[ ] can be found at [ ]," describe a picture with sentences, or write a story with collaboration with other users. OpenMind website was launched in 1999 and gathered more than 700,000 common-sense sentences for five years until March of 2006.

Because the OpenMind website already had more than 700,000 common-sense sentences which had been gathered for five years, I decided to reuse the OpenMind common-sense knowledge rather than to start from the scratch. However, since the OpenMind common-sense knowledge was focused on English sentences without any cultural information, we still needed to gather more information from various cultural backgrounds written in various languages.

To gather multilingual/multicultural common-sense knowledge, I built and launched GlobalMind website. GlobalMind website [16] is designed to gather common-sense knowledge data from various cultures and various languages as well as relationships and connections between common-sense of different cultures. The basic structure of the website is almost the same as the structure of the OpenMind website. Users can type in their common-sense knowledge by filling in blanks in templates. They can choose their own languages to use among various languages the website supports. Additionally GlobalMind supports bilingual/bicultural activities to gather connections between different language/cultures. Users can read a sentence written by other users in different cultural backgrounds and translate the sentence to their own languages or evaluate the strength of the common sense in their own culture/language. Figure 4-2 and Figure 4-3 show examples of monolingual and bilingual activities in GlobalMind website.

To encourage the participation of volunteers, we showed the statistics of collected items at Front Page sorted by number of contribution. We also held a gift-certificate event for ten weeks, giving Amazon, Amazon Japan, or Interpark Korea gift certificate to the two best teachers each the week.

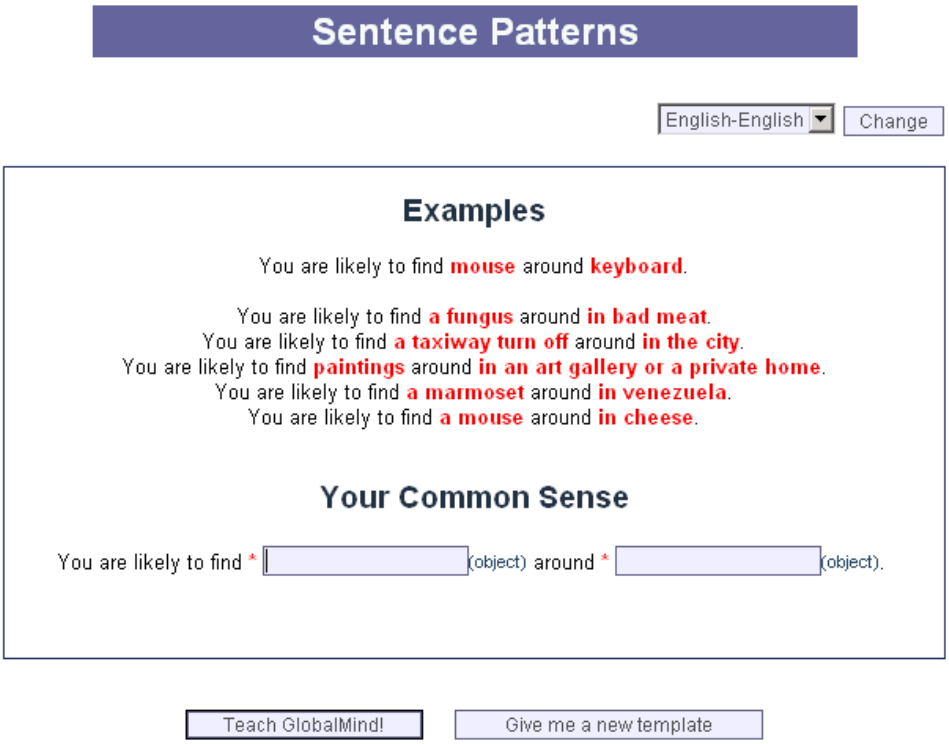


Figure 4-2: Monolingual activity of GlobalMind website

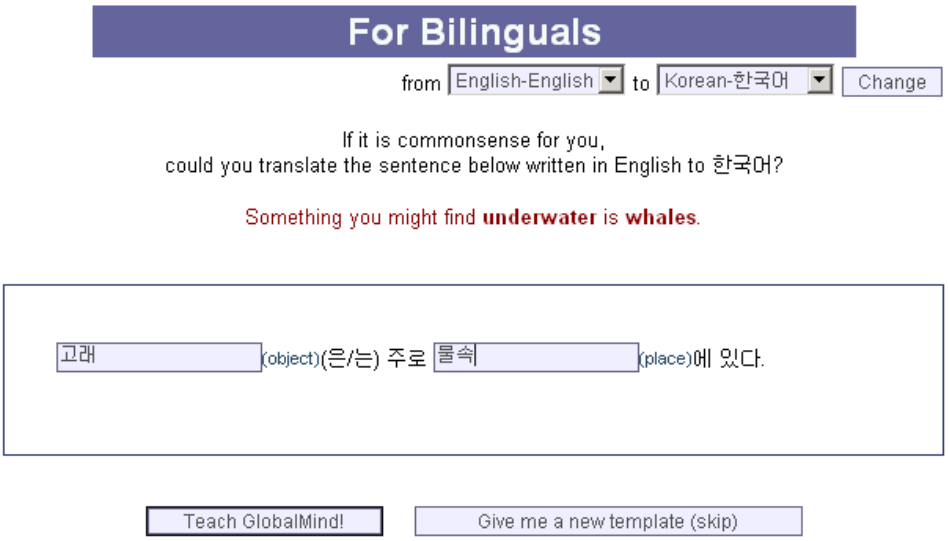


Figure 4-3: Bilingual activity of GlobalMind website

GlobalMind website is launched December 12, 2005 with four languages including English, Korean, Japanese, and Chinese with both of Simplified and Traditional Chinese. As the date of June 14, 2006, GlobalMind website has gathered 32254 common-sense sentences excluding data from original OpenMind, and 11023 bilingual/bicultural connections. Table 4.1 shows how many items and data have been accumulated by GlobalMind as the date of June 14, 2006. The table excludes data from original OpenMind.

Table 4.1: Statistics of data accumulated through the GlobalMind website

Languages	
Korean	15140
Japanese	9010
English	7787
Chinese	317
total	32254
Cultural Backgrounds	
Korea	19031
Japan	9129
Germany	1657
USA	1360
Finland	212
Taiwan	208
Unknown	190
Malaysia	160
Australia	154
etc	153
total	32254
Bilingual Connections between	
English and Korean	5556
English and Japanese	4444
Japanese and Korean	733
Chinese and Korean	218
Chinese and Japanese	58
Chinese and English	9
Chineses	2
total	11023

4.3 Processing of Accumulated Data

Common-sense knowledge sentences accumulated through the GlobalMind website and the OpenMind website should be processed to form a GlobalMind network. Each sentence is transformed into a pair of two nodes and a link between the nodes. For example, common-sense sentence "a student can be found at a school" is transformed to two nodes "student" and "school" and a link "LocatedAt."

The GlobalMind website processes and transforms the sentences at the time when the sentences are typed in. This approach has some advantages and disadvantages. It can reduce processing time and resources and provide a database that is updated in real time. On the other hand, it cannot make use of more professional natural-language processing because it requires the processing of the sentences in a relatively short time. To compensate for this disadvantage, we store both of processed predicates and unprocessed sentences; if we need to use more professional natural-language processing then we can load unprocessed data and process them at any time.

To reduce the processing resources and time, GlobalMind uses a simple template matching system. If a user entered "a student can be found at a school," and GlobalMind has a template "[ ] can be found at [ ] / LocatedAt," then GlobalMind strips off the template from the sentence and make it a pair of two nodes and a link like "[a student] [a school] [LocatedAt]." This is a predicate in GloalMind and represented by "a student–>LocatedAt–>a school."

We still need to lemmatize the predicate because otherwise a predicate "a student–>LocatedAt–>a school" is regarded as different from a predicate "students–>LocatedAt–>schools" that has the same meaning from the former. In the case of English phrases, GlobalMind website uses MontyLingua to lemmatize phrases [30]. GlobalMind tokenizes a node, filters out stop words such as "a" or "the", and lemmatize each token with Liu's lemmatizing tool. Lemmatized tokens are merged into one node again. For example, a node "the car's battery dies" is transformed into "car battery die."

In the case of Korean and Japanese, which are more complicated than English to lemmatize, GlobalMind depends on the template structures. The templates for Korean and Japanese are designed to encourage users to enter only the stem parts of words. This method is chosen because it takes much less time and computing resources than using other Korean or Japanese natural-language processors. GlobalMind also saves the raw sentences of Korean and Japanese data so they can be processed by more professional natural language processing tools when necessary.

4.4 Global Network Building

The accumulated common-sense predicates forms the networks of GlobalMind. However, we still need to connect the networks with each other to build the large network. In connecting networks GlobalMind makes link-to-link connections rather than node-to-node connections.

In this step, GlobalMind is expected to make basic connections which should be sufficient to make reasonable inferences among different networks, but do not cover the entire network.

GlobalMind depends on bilingual or multilingual volunteers to make these basic connections between different networks. The bilingual/multilingual activities introduced in Section 4.2 are used for these basic connections. For example, GlobalMind bilingual activities ask bilingual users to translate a sentence into their native languages if the sentence is common sense in their cultures. This translation is not for gathering new common-sense knowledge, but for connecting existing knowledge in two different networks. Thus, translated sentences are stored with translated tags to be distinguished from common-sense knowledge gathered in the native language. When a sentence in one language is translated into another sentence in another language, GlobalMind considers the connection between these two languages is established.

With these basic connections established, the Inference Module, described in the next sec-

tion, will operate to correlate the networks.

4.5 Similar-concept Inference Module

In this section, I will describe the Similar-concept Inference Module (SIM) and how it works. SIM is a module which finds the most similar concepts in a target language/culture when there is a given concept in one language/culture. For example, if a user gives "forks and knives" to SIM in English/America and asks the similar concepts in Korean/Korea, SIM returns "SUD GA RAK, JEOT GA RAK(spoons and chopsticks)" because they are both the main utensils for main meals in each culture.

SIM uses the expand-and-contract method. SIM provide two kinds of inference - finding similar nodes and finding similar links. Because both of these two inferences are basically using the similar expand-and-contract method, in this section I will explain the method with finding similar nodes and then expand the method to finding similar links.

As described in Section 3.2.1, the expand-and-contract method is to find a node with similar concepts to a given node when two nodes are in different networks and written in different languages by comparing topologies of each network. When there is a given node, SIM starts extracting related common-sense about the given node, which means a sub-network around the given node. With the extracted sub-network, SIM tries to find the most similar sub-network in a target network, and then contract the target sub-network into one node, which is a target node, by comparing the topologies of a given sub-network and a target sub-network.

4.5.1 Expanding the Sub-network

The input of SIM is a concept, a language/culture, and a target language/culture. When SIM reads the input data, the first task SIM does is expanding, which means extracting the

sub-network around the given concept.

Before expanding the sub-network, SIM should decide how deep and broad of a network we would use for the comparison. Let me define several terms here. When there is a given node, I called it as a root node, and a sub-network with the root node only is Level 0 network. The root node will be connected with other nodes via links, and the other nodes are called as children nodes and the links between the children nodes and the root node are called as children links. The sub-network of the root, the children nodes, and the children links is Level 1 network. In the same way, a sub-network can be expanded to Level 2 with grandchildren nodes and links and to Level 3 with great grandchildren nodes and links.

4.5.2 Finding the Matching Sub-network

After expanding the sub-network in a given network, the next task is finding the matching sub-network in a target network. This task is done based on the bilingual connections between two language/culture networks.

SIM finds bilingual connections between the given sub-network and the target network. As already described, bilingual connections mean two predicates each of which is located in each network and both of which have similar meanings.

Because the given sub-network is Level 3 sub-network from the given node, the bilingually connected predicates in the given sub-network should be within the distance of three levels from the given node. Thus, we can assume that the target node is also within the distance of three levels from the bilingually connected predicate in the target network. From the assumption, SIM extracts target sub-networks which are Level 3 sub-networks from the bilingually connected predicates. Because there can be several bilingual connections in the sub-networks, the final target sub-network is the union of all the target sub-networks extracted.

4.5.3 Contracting the Sub-network

After finding the matching sub-network, we now have a pair of sub-networks: a given sub-network and a target sub-network. Contracting is a process to find a target node with the most similar concepts to the given concept in the target sub-network by comparing two sub-networks. In this step, we compare the topology of two sub-networks, score each node with the topology structures, and find a target node with the biggest score.

Basically SIM compares the routes; if a node in a target network has the same routes the given node has, SIM adds a score to the node. The scoring system is described below, but let me show a simple example first. If a given node "school" has a route to a node "child" via "school<–LocatedAt<–student–>IsA–>child," a node in a target network "HAK KYO(school)" has a route to a node "EO RIN YI(child)" via "HAK KYO(school)<–LocatedAt<–HAK SAENG(student)–>IsA–>EO RIN YI(child)," and there is a bilingual connections between "child" and "EO RIN YI(child)," which means "EO RIN YI(child)" might have the same/similar concept to "child," then the possibility that "HAK KYO(school)" has the same/similar concept to "school" is higher than when they don't have the same routes. Thus, when SIM find the same or similar routes between nodes and the root nodes in both sub-networks, SIM adds score to the nodes. After all the scoring, SIM sorts the nodes by the scores and shows two candidate nodes with the highest scores.

In this network topology comparison there are several factors we should consider, such as kinds of relationships, number of children nodes, and the distance between nodes. Here the factors are represented as weights of links.

Weight of Links

In sub-networks, all the nodes are connected with other nodes forming a network. However, the importance and the strength of the connections can be all different. Thus, we should consider how to weight the links by their importance and the strength. In the GlobalMind

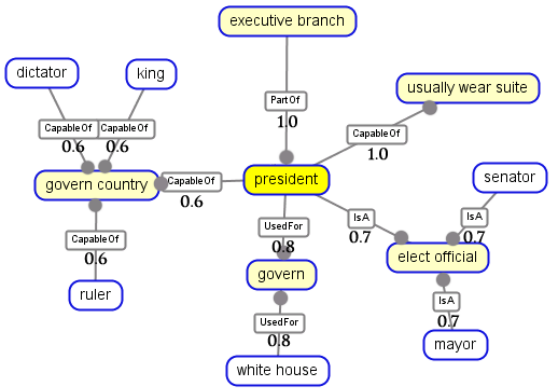
SIM weight system, all the links have their own weight between 0 to 1, where it is 0 if the connection is weak and 1 if the connection is strong, and each weight is calculated by three factors below. When SIM scores nodes, it uses the weights of the links on the routes from the root node to the nodes.

The first factor considered is a number of children nodes of each node. In Liu's ConceptNet system, the strength of link is affected by the number of children nodes [26]. According to Liu, connection between two nodes becomes weakened as the nodes have more number of children. For example, a node "heat" and one of its twelve children nodes "CapableOf-cause fire" have a stronger connection than a node "person" and one of its 3000 children nodes "CapableOf-build." This theory is also adpated to GlobalMind SIM.

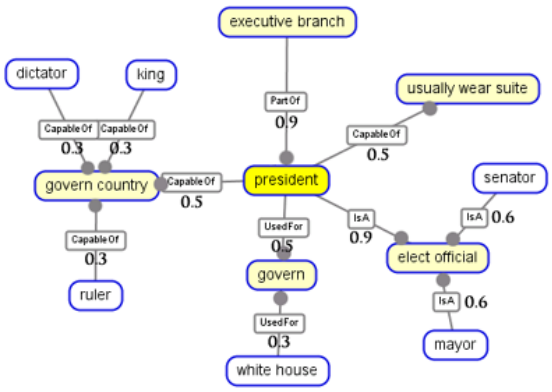
The second factor in the weight system is a distance from the root node. It is obvious that a child node is more related with a root node than a grandchild node, because the grandchild node is related with the root node through the child node's relationship with the root node. The weight from the distance is automatically applied by combinations with other factors. The process will be described at the end of this chapter.

Another factor, which can be considered but not implemented in this thesis, is the kind of relationships. As described in Section 3.1.2, all the nodes in GlobalMind are connected with other nodes with 22 different kinds of relationships, some of which have strong connections and others of which don't. Thus, we should consider the kind of relationships between nodes. For example, two nodes "apple" and "fruit" which are connected with the "IsA" relationship might have a stronger connection than other two nodes "dog" and "run", which are connected with the "DesireOf" relationship.

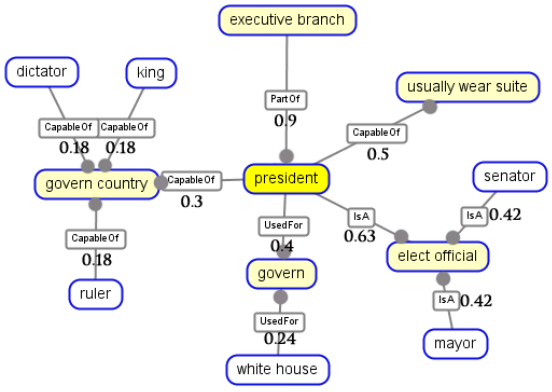
The weight of each link can be calculated with the combination of the weighted factors written above. Let me explain the weight calculation process with Figure 4-4. The weights and the numbers of Figure 4-4 are not the real weights and numbers used in GlobalMind SIM but chosen to show the obvious effects of weights.



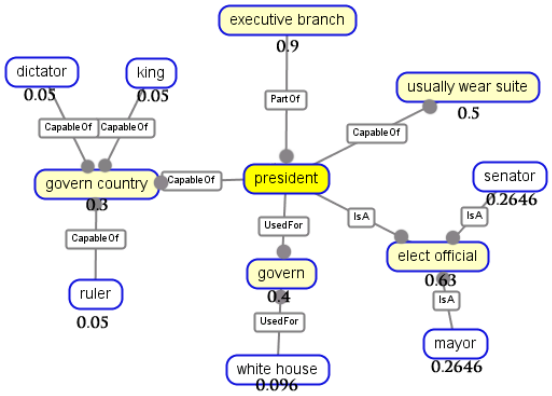
(a) Weights by the number of children



(b) Weights by relationship



(c) Weights by relationship and the number of children



(d) Weighted sub-network

Figure 4-4: Weights of sub-networks. The weights in the figures are not the real numbers used in GlobalMind

Figure 4-4(a) shows the weighted links with a number of children nodes. In Figure 4-4(b), you can see the links weighted with relationship-related weight factors. Each different relationship link has different weight. These two kinds of different weights are multiplied with each other, and the result is shown in Figure 4-4(c).

At this point, we have the weights of links between any two directly-connected nodes. However, we should consider the importance of relation between two nodes which are not directly connected but connected through a few other nodes and links. Thus, here is the process to compute the importance of relation between any two nodes.

When there are two nodes, there are nodes and links between them. The order of nodes and links on the way from one node to the other node is called a route. For example, in Figure 4-4, "president–>CapableOf–>govern country<–CapableOf<–ruler" is a route between a node "president" and a node "ruler." The importance of the relation between two nodes can be represented by a weight of the route between two nodes. For an easy description, let me explain the process of weighting of the routes between all the nodes and the root node.

At first, we already have an appropriately weighted Level 1 sub-network. The weighted Level 2 sub-network can be computed from Level 1 sub-network. A grandchild node has a route to the root node via a child node, and the grandchild route's weight is recalculated by multiplying the weight of the grandchild link and the weight of the child route. In Figure 4-4(c), the weight of the route "president–>IsA–>elect(ed) official<–IsA<–senator" will be recomputed to 0.2646 which is the multiplied number of 0.63 and 0.42. The same method is recursively used in Level 3 sub-network or more level sub-networks. Because the weights are between 0 and 1, a weight is decreased, or at least not increased, as a distance between a node and the root node is increased. Figure 4-4(d) shows the final version of the weighted sub-network.

In the real process of comparing and scoring, SIM does not know what will be the root node in a target network. In fact, all the processes described here are for finding the root

node. Thus, SIM does not pre-calculate all the weights of routes before comparing. Rather, comparing and scoring processes dynamically calculate the weights of routes between two nodes when necessary.

Comparing and Scoring

At this point, we have weighted sub-networks, one of which is a given sub-network and the other is a target sub-network. In this section I will explain how SIM compares two sub-networks, scores nodes, and finds target nodes we want.

The comparison starts from the bilingual connections established by bilingual activities described in Section 4.2. At first, SIM searches the bilingual connections between two sub-networks.

In this process SIM works with two assumptions. The first one is that when two nodes have the same or similar routes from one node, these two nodes might have similar concepts. Here, the route means the orders, the relationships, and the directions of links. Extension of this assumption is that, if two descendant nodes have similar routes from other two ancestor nodes, and these two ancestor nodes have similar meanings to each other, then those two descendant nodes might have similar meanings too. The second assumption is that if there is a bilingual connection established between two predicates in two different networks these two predicates might have similar concepts or meanings, and thus two pairs of nodes have similar concepts in the context of the links in the predicates.

From the assumptions, SIM tries to find a target node which has a route from a bilingually connected node and the route is similar to the route between a given node and the bilingually connected node in a given network. If a node #G1 and a node #G2 are in a given network, if a node #T1 and a node #T2 are in a target network, if there is a bilingual connection between the node #G1 and the node #T1, and if the route #G between the node #G1 and the node #G2 and the route #T between the node #T1 and the node

#T2 are same or similar, with all the conditions altogether SIM regards a node #G2 and a node #T2 having similar meanings. Thus, if the node #G2 is a given node, the node #T2 becomes a target node.

Thus, here SIM considers two factors. The first one is the similarity of routes in a given network and routes in a target network. The second one is the importance of routes which is calculated in the previous section.

At first, SIM finds bilingually connected nodes in sub-networks. From the nodes, SIM compare route #Gs and route #Ts. If a route #G is similar to a route #T in a meaning of the order, the relationship, and the direction. If two routes are similar, SIM increases a score of the node #T2 by the weight of the route #T between the node #T1 and the node #T2. Thus, the higher score means the higher possibility to be a target node.

After the comparing and scoring process, SIM regards that the node with the highest score is the target node. Currently, SIM shows to users top two nodes with highest scores as the first candidate target node and the second candidate target node. The evaluation results of SIM process will be discussed in Section 5.1.

4.5.4 Inference on Similar Links

Up to now, I have described how SIM finds the target node with the most similar concept to the given node. SIM also provides finding the target predicate with the most similar concept to the given predicate. Basic method is same to the method used for the target node finding. SIM regards a given predicate as a kind of node during the inference process and executes the similar processes to find target nodes. After finding several candidates of target nodes, SIM considers the links between nodes and finds the target predicates.

4.6 Difference Inference Module

In this section, I will describe the Difference Inference Module (DIM) and how it works. DIM is a module which finds the cultural differences between a given language/culture and a target language/culture when there is a given situation. For example, if a user gives "restaurant" to DIM and asks the cultural differences between English/America and Korean/Korea, DIM returns "in America you should give waiters tips / in Korea you don't give waiters tips" or "in America there is one waiter for your table / in Korea every waiter serves every table."

Inputs of DIM are usually two networks and one situation. Two networks are two language/cultures to compare the differences. The situation can be one node like "restaurant" or a combination of several nodes like "restaurant, evening, and birthday."

Basically DIM extracts sub-networks related to a give situation from both networks, compare the sub-networks with each other, removes the shared common-sense predicate, and then returns the left common-sense unit which might be differences between two networks.

DIM can be generalized to and used for any pair of language/culture. In this section, however, I will handle only the case of "English/America" and "Korean/Korea" and then generalize it at the end of the chapter to reduce the complexity of description.

4.6.1 Situation Analysis

DIM starts from analyzing given inputs such as situations.

Given a situation, the situation can be a word like "restaurant" or a combination of several order-dependent words like "restaurant, evening, and birthday." DIM regards each word as a node. Thus, if the situation consists of N words, the situation is analyzed as a combination of N nodes. Although all the nodes in the situation will be concerned in the inference

process, the first node of the situation is considered as the most important situation while the last node is considered as the least important situation. In the example of "restaurant, evening, and birthday," DIM will try to find the cultural differences about restaurants, in evening, on a birthday, but focusing on restaurants rather than a birthday. Thus if there is no information about "restaurants on a birthday," DIM will return information about "restaurants."

4.6.2 Sub-network Extraction

After analyzing the situations, the next step in DIM is the extracting of sub-networks.

The situation is usually written only in one language. Here I will assume that the given situation is written in English.

When the situation is nodes written in English, we can easily extract the American sub-network by extracting Level 1 sub-networks with root nodes which are same or similar to the situation nodes. Because there could be several situation nodes, and also one situation word can be represented by several nodes, the extracted sub-network may be a combination of several Level 1 sub-networks.

Not only from the American network, but also from the Korean network should we extract the sub-network with a given situation while the situation is written in English. Thus here we need to translate the situation into Korean. GlobalMind DIM is using online machine translators to translate the situations and other data. Currently DIM is using Google machine translator [17]. After translating the situation into Korean, the Korean sub-network is extracted by the same way by which American sub-network is extracted.

Level of sub-network to be extracted can be discussed in further research. However, in this thesis, we only use Level 1 sub-network because even Level 2 sub-networks included too many information that are not strongly related with the situation. In the case of SIM,

which contracts the result into the most relevant nodes before it returns results, we can use information as many as the computer program can handle. However, DIM does not have the contracting process. Thus we need to prune irrelevant information from the first step of the inference processes.

4.6.3 Comparison and Removal

Now we have two sub-networks, each of which is from each network. DIM compares the sub-networks with each other sub-networks, removes the same or similar common sense, and returns the left sub-networks which means the differences between two networks.

How can we find the shared common sense? At first, if there are bilingual connections between two predicates in the two sub-networks, then they are the shared common sense and should be removed. DIM finds the bilingual connections between two sub-networks. Considering that the bilingual connections are a kind of translation, and the translated predicates are not regarded as original common sense in the network, removing the bilingual connections itself is nothing but removing the connections. The predicates which should be removed are not the translated predicates themselves but the predicates which are original common sense in the network and similar to the translated predicates at the same time.

As already described, there are not so many bilingual connections compared to the number of predicates. Thus, using this method alone is not enough. We need another method to improve the comparison.

If these two sub-networks are written in the same language, English, we can simply find the shared common sense by comparing the text of each predicate. If American sub-network has a predicate "student–>LocatedAt–>school" and Korean sub-network also has a predicate "student–>LocatedAt–>school," this can be regarded as the shared common sense and can be removed. Thus, if DIM translates Korean sub-network into English, the language of American sub-network, it can easily compare and find the shared common sense.

DIM uses a Google web machine translator [17] to translate Korean sub-network into English. After translating the Korean sub-network into English, DIM compares each predicates with the predicates in American sub-network. Because the Korean-English machine translator is not good enough, we cannot expect the texts of two predicates will be exactly matched when they have the same meanings. Rather DIM regards them as the same or similar predicates when they have same words in them. For example, if a predicate A consists of a node A1, a node A2, and a link A between a node A1 and a node A2, and the other predicate B consists of a node B1, a node B2, and a link B, the predicate A and the predicate B are regarded as the shared common sense when a node A1 and a node B1 contains at least one same word, a node A2 and a node B2 contains at least one same word, and the link A and the link B have the same relationship and the same direction. Here "the same words" does not mean that two words are exactly matched by character by character, but means that two words have the same word stems. Also, prepositions such as "on" and "with" and stop words such as "the" and "a" are not included in this comparison.

Table 4.2 shows that how Korean predicate A is translated into English. As the table shows, the machine translator does not provide decent translation. Thus, DIM compares the words in each node. In the table, the underlined word "wedding" is matched in node 1s and the other underlined word "dress" is matched in node 2s. Because there are at least one matched word in each node and the links are the same, the predicate A and B are the shared common sense.

Table 4.2: Comparison of two predicates in different languages

Predicate A	
original	GYUL HON SIK –> OnEvent –> WE DING D RE S RUL IB DA
original meaning	wedding –> OnEvent –> wear wedding dress
machine translated	wedding ceremony –> OnEvent –> the [wey] [ting] puts on the dress
Predicate B	wedding –> OnEvent –> wear wedding dress
Shared Word	wedding –> OnEvent –> dress

After removing all the shared common sense, the left sub-networks are returned as the cultural differences between two networks. The quality of the left sub-networks as the

cultural differences is dependant on the quantity and quality of both of Korean/Korea and English/American common-sense database. At this point, because of limited amount of Korean common sense, many of the American common-sense assertions which are also true in Korea are not removed by Korean GlobalMind database and returned as the differences. However, I hope this problem will be resolved as the database is enlarged.

Chapter 5

Evaluation

The GlobalMind project consists of two major factors: databases and inference modules. While the statistics of data accumulated in the databases are described in Section 4.2, this chapter describes the processes and results of evaluation for inference modules.

The performance of inference modules is mostly dependent on the quality and quantity of databases. At this point, the size of GlobalMind databases is not large enough to make perfect inference. Thus, this evaluation is aimed to test the potential and to search for the future direction of improvement of GlobalMind rather than to prove the performance of inference modules.

Because the English database is the largest database, and the Korean database is the second largest database in GlobalMind, this evaluation is done with English and Korean databases.

5.1 Similar-concept Inference Module

GlobalMind Similar-concept Inference Module extracts the concepts which are similar to a given concept. The extracted similar concepts can be dictionary words for the given con-

cept, or they can be different from dictionary words but have similar concepts from the given concept.

This evaluation is designed to test if SIM can extract the similar concepts in relatively high probability, and if SIM can extract the concepts which are similar to a given concepts but cannot be found in a dictionary. Because of the limited size of databases, we cannot expect the best result. However, this evaluation can determine whether there is potential in GlobalMind SIM or not.

5.1.1 Design

GlobalMind SIM is given English concepts and extracts the most similar Korean concepts for the given words. The similarity of a given English word and an extracted Korean word is measured.

Korean human subjects evaluate whether the words in each pair have the similar concepts or not. Each pair will be divided into four categories: if the English word and the Korean word share the same dictionary meaning, "same," if they do not have the same meaning but are conceptually similar based on contexts, "similar," if they are neither same nor similar but if the subject automatically reminds the other word when s/he sees/hears one word in the pair, "related," and in other cases, "not related at all." For the example of "fork" in English/America, "PO K (fork)" is "same," "JEOT GA RAK (chopsticks)" is "similar," "SIK SA (meal)" is "related," and "NAM JA (man)" is "not related at all."

In "same" and "similar" pairs, the English word and the Korean word can substitute each other, while in "related" and "not related at all" pairs, cannot. Thus, for simple comparison, "same" and "similar" can be grouped as "matched," and "related" and "not related at all" can form another group, "unmatched." In the best case, all the pairs will be evaluated as "matched." In the worst case, all the pairs will be evaluated as "unmatched."

The word pairs are also evaluated with an English-Korean dictionary. The English word in a pair will be looked up in the dictionary, and whether the Korean word in the pair can be found in the dictionary or not is checked. Because the goals of a dictionary and SIM are different, the rate of this test does not directly show the performance of SIM. This result is compared to the result from human subjects to analyze how GlobalMind can find similar but not the same concepts.

Test Concept Sets

The given concepts were chosen from the 300 most frequently used English words [9]. Among the 300 words, the words whose primary meanings are nouns were chosen, and the other words such as "a," "and," "to," and "also" were removed. After the removal, 72 English nouns were given to SIM. GlobalMind SIM extracted the most similar Korean concepts for 61 English words among 72 words, while 11 words couldn't be processed. Table 5.1 shows the 61 words SIM processed. These 61 words are used for the evaluation.

Table 5.1: 61 English words processed by SIM

air	America	animal	book	boy	car	children
city	country	day	earth	example	eye	face
family	father	feet(foot)	food	girl	hand	head
help	home	house	Indian	land	letter	life
light	line	man	mother	mountain	night	number
oil	page	paper	people	picture	place	plant
point	read	river	school	sea	sentence	side
song	sound	story	study	thought	time	tree
watch	water	white	word	world		

Normally SIM generates two candidate concepts for each English word: the first candidate and the supplementing second candidate. If SIM cannot find the second candidate, it only provides the first candidate. In this evaluation SIM found two candidates for 52 English words and one candidate for the rest nine words. Total 113 pairs are generated: 61 first-candidate pairs and 52 second-candidate pairs. In this evaluation, the first candidate pairs and the second candidate pairs will be separately treated.

Human Subjects

Human subjects must be very familiar with Korean culture and Korean language, and be able to read and write in English. Korean people who have lived in Korean more than 20 years are chosen as human subjects. Seven Korean people including five males and two females participated. Ages are between 24 and 29 where the average is 26.5, and the durations of living in Korea are between 20 and 28 where average is 24.

Evaluation Form

Participators are asked to fill out an on-line evaluation form. The form shows pairs of an English word and a Korean word. The subjects choose the relationship between the two words among "same," "similar," "related," and "not related at all." Figure 5-1 shows the on-line evaluation form used in the evaluation.

5.1.2 Result

As described above, SIM extracts two candidate concepts for one given word. Because the first candidate is the main result and the second candidate is supplementary result, I will analyze the first candidates and the second candidates separately. The whole result of the survey can be found in Appendix A.

Table 5.2 shows the answers of subjects assorted by the class of candidates. Count means how many times each answer selected by human subjects. If a person selects "same" for 30 pairs and "similar" for 10 pairs and another person selects "same" for 25 pairs and "similar" for 5 pairs, the count of "same" is 55 and the count of "similar" is 15. Because there are 61 first candidate pairs and seven human subjects, the maximum count of answers for the first candidate pairs is 427.

With four choices of "same," "similar," "related," and "not related at all," the probability of each answers in random selection is 25%. As described in Section 5.1.1, "same" and "sim-

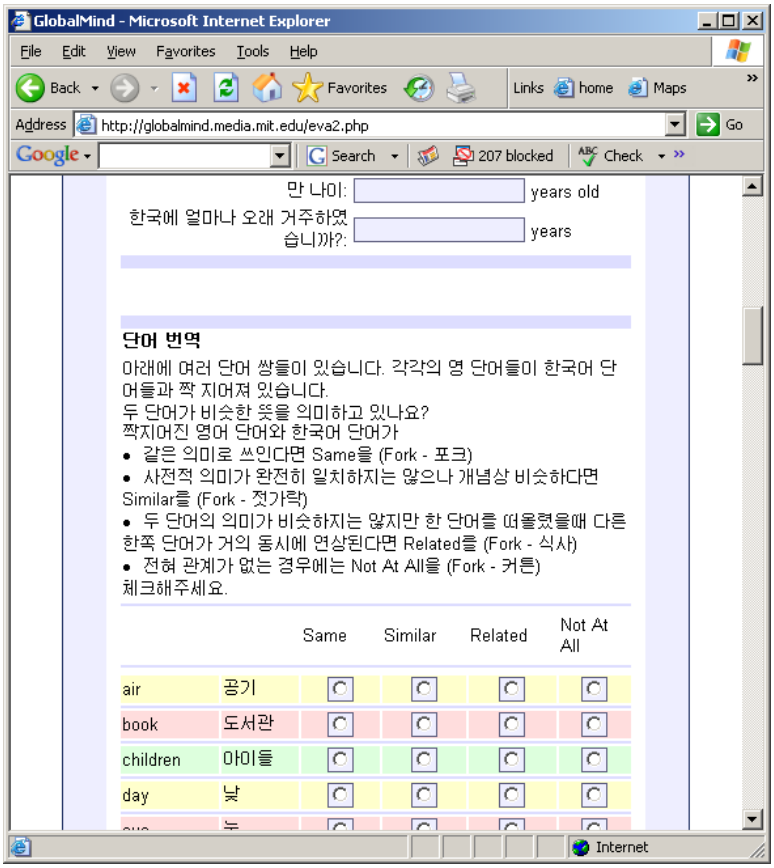


Figure 5-1: SIM evaluation survey forms

Table 5.2: Human answers for SIM evaluation

	relationship	count	rate
First Candidate	same	325	76.47%
	similar	39	9.18%
	related	35	8.24%
	not related at all	26	6.12%
	matched	364	85.65%
	unmatched	61	14.35%
Second Candidate	same	135	36.39%
	similar	39	10.51%
	related	106	28.57%
	not related at all	91	24.53%
	matched	174	46.90%
	unmatched	197	53.10%
Total	same	460	57.79%
	similar	78	9.80%
	related	141	17.71%
	not related at all	117	14.70%
	matched	538	67.59%
	unmatched	258	32.41%

ilar" can be grouped as "matched" and "related" and "not related at all" can be grouped as "unmatched." Thus, the rate of each group is 50% in random selection. Because the goal of GlobalMind SIM is searching the words with the same or similar concepts, if it is working, the rates of "matched" become high and that of "unmatched" become low. Thus, if the result of SIM test shows the rate higher than 50% in "matched" and the rate lower than 50% in "unmatched," we can say SIM is working as it was intended to.

In the case of the first-candidate pairs, the result shows that the 76.47% of first-candidate pairs have the same concepts, and the 9.18% pairs have the similar concepts. The "matched" first-candidate pairs are 85.65% of the whole first-candidate pairs which is higher than 50%. The rate of "unmatched" first-candidate pairs is 14.35% which is lower than 50%. With the null hypothesis of that the rate of the "matched" first-candidate pairs will be lower than 50%, and the alternative hypothesis of that the rate of "matched" first-candidate pairs will be higher than that of "unmatched" first-candidate pairs, the p-value is less than 0.001, which means there is high likelihood that the alternative hypothesis is true. It shows that

the first candidates are not perfect but well inferred and meaningful.

Considering the small-size databases which limit the performance of SIM, it shows the strong potential of inference algorithms. In the most case of "unmatched" first-candidate pairs, GlobalMind Korean database itself does not have the matching word for the given English words at all. Thus, we can guess that the main reason of SIM's failure is the limited size of database rather than the failure of inference algorithms, and the performance of SIM can be improved by adding more common-sense knowledge into databases.

In the case of the second-candidate pairs, 46.90% of the pairs are "matched" and 53.10% of the pairs are "unmatched." Although the rate of the "same" second-candidate pairs are 36.39% which is higher than 25%, the expected rate in random selection, the rates of "related" and "not related at all" are also 28.57% and 24.53% which are higher than or close to 25%. With the same null hypothesis and the alternative hypothesis as the case of the first-candidate pairs, the p-value of the second-candidate pairs is 0.884. Thus, we can conclude that the supplementary candidates are not as meaningful as the first-candidate pairs are.

The pairs generated by GlobalMind SIM was also compared to the Yahoo English-Korean dictionary [45]. If the Korean word in a pair can be found when the English word in the pair is looked up in the dictionary, the pair is marked as "confirmed," and in the other case, "unconfirmed." The 49 first-candidate pairs out of the 61 first-candidate pairs are "confirmed," and the 20 second-candidate pairs out of the 52 second-candidate pairs are "confirmed" by the dictionary. "Unconfirmed" pairs include wrong inferences and indirect inferences.

Here my hypothesis is that GlobalMind SIM can find the similar concepts which are different from dictionary words but have the same uses based on contexts. If SIM can only

find the words in a dictionary and cannot make inference based on contexts, "matched" pairs and "confirmed" pairs will be the same and "unmatched" pairs and "unconfirmed" pairs will be the same. If the hypothesis is correct, some of "unconfirmed" pairs will be "matched" pairs, mostly "similar" pairs, and the rest of "unconfirmed" pairs will be wrong inferences. If the hypothesis is not correct, all "unconfirmed" pairs will be wrong inferences and there will be no "matched" pairs among "unconfirmed" pairs.

Table 5.3: Human answers on unconfirmed pairs

	same	similar	related	not at all	matched	unmatched
First Candidate						
Total Answers	10	14	33	26	24	59
rate	12.05%	16.87%	39.76%	31.33%	28.92%	71.08%
Second Candidate						
Total Answers	16	23	99	86	39	185
rate	7.14%	10.27%	44.20%	38.39%	17.41%	82.59%
Total Answers	26	37	132	112	63	244
rate	8.47%	12.05%	43.00%	36.48%	20.52%	79.48%

Table 5.3 shows the result of people's answers to the "unconfirmed" pairs. 28.92% of the "unconfirmed" first-candidate pairs and 17.51% of the "unconfirmed" second-candidate pairs are "matched." The rest of the pairs are "unmatched" pairs which means wrong inferences. If the hypothesis was incorrect, the rate would be 0%. Thus, here we can find SIM can find the matching words which are missed in a dictionary.

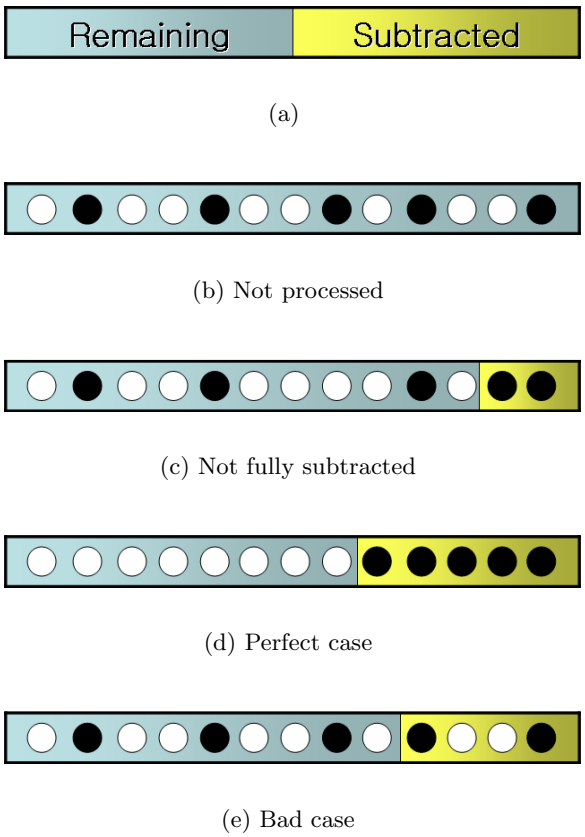
5.2 Cultural Differences Inference Module

GlobalMind Differences Inference Module extracts cultural differences about specific topics. To infer the cultural differences between two cultures, at first DIM extracts all common-sense knowledge related to a given topic, and subtracts common-sense knowledge that are shared by both cultures. The remaining common-sense knowledge after subtraction is cultural differences DIM provides.

The performance of DIM is largely influenced by the subtraction. Two factors are important in the quality of subtraction: the quality of subtracted common sense and the

number of subtracted common sense. At first, DIM should subtract only shared common-sense knowledge; if DIM subtract not-shared common sense by mistakes, the performance will be lowered. Secondly, DIM should subtract shared common-sense knowledge as much as possible; if DIM cannot subtract much of shared common-sense knowledge, the suggested cultural differences will include knowledge that are not "differences."

Figure 5-2 shows the concept of this process. Each circle represents each common-sense knowledge where black circles are shared common sense and white circles are different common sense. Figure 5-2(b) shows the initial knowledge set that are not processed yet. In ideal case, as subtracting the set temporarily looks like Figure 5-2(c) and finally looks like Figure 5-2(d). In the bad case, it can subtract not-shared common sense by mistakes and it will look like Figure 5-2(e).



(a)

(b) Not processed

(c) Not fully subtracted

(d) Perfect case

(e) Bad case

Figure 5-2: GlobalMind DIM processes and performance

DIM determines the shared knowledge by data accumulated in the databases; if a knowledge is located in both of American and Korean databases, it is a shared knowledge. Thus, the number of shared knowledge is mostly dependent on the size of databases while the quality of subtracted knowledge is mostly dependent on the inference module. With the limited databases, this evaluation will not measure the number of subtracted shared knowledge but measure the quality of subtracted shared knowledge. In Figure 5-2, this evaluation will measure to which DIM is closer between Figure 5-2(c) or Figure 5-2(e) rather than between Figure 5-2(c) and 5-2(d).

5.2.1 Design

For given situations, GlobalMind DIM makes inference on cultural differences between America and Korea. During the process DIM generates the initial set, the subtracted set, and the remaining set. The initial set is the collection of all common-sense knowledge related to the given situations, the subtracted set is the collection of all shared common-sense knowledge determined by DIM, and the remaining set is the cultural differences provided by DIM.

In this evaluation, the proportion of "differences" and "similarities" of each set is tested. If the inference algorithm works as it is intended to, the proportion of "differences" of the initial set will be higher than that of the subtracted set and lower than that of the remaining set. In the best case, the proportion of "differences" of the subtracted set is 0%, that of the remaining set is close to 100%, and that of the initial set is between them. In the worst case, the proportions of all sets will be the same.

The "differences" and "similarities" are evaluated by human subjects. American human subjects and Korean human subjects evaluate each knowledge sentence if it is common sense in their own cultures or not. If both Korean and American agree the sentence is common sense in their cultures, the sentence is one of "similarities," and if not, "differences."

Test Common Sense Sets

Two topics, funerals and restaurant, are chosen for this evaluation because both topics are familiar with people and both topics have some knowledge in the database. The English commonsense knowledge related with funerals and restaurant is 63 sentences including 13 sentences about funerals and 50 sentences about restaurant. The whole sentences used in this evaluation is attached in Appendix A.

DIM processed the initial set and made the remaining set with 37 sentences and the subtracted set with 20 sentences. Table A.3 shows the remaining set, and Table A.4 shows the subtracted set.

Human Subjects

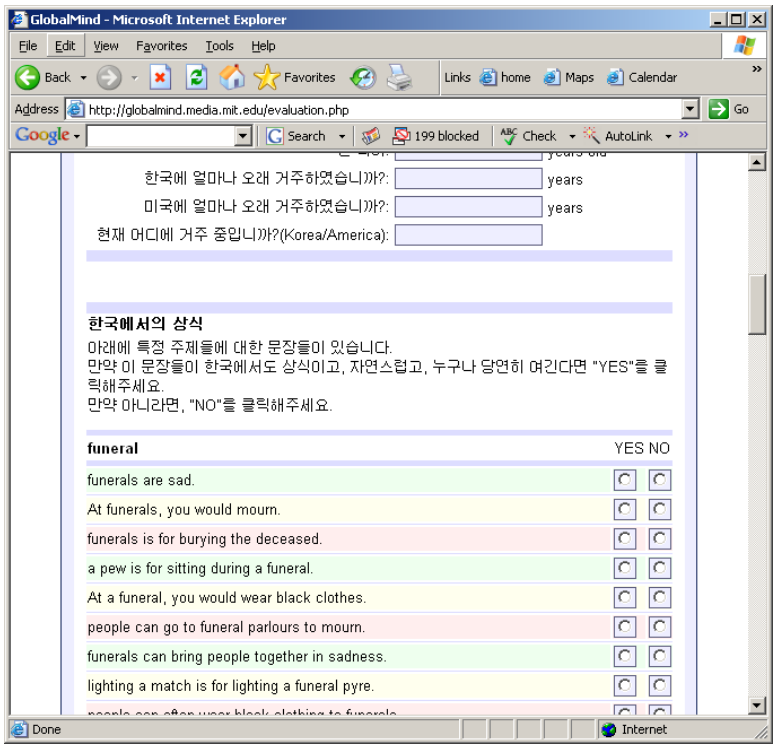
Korean human subjects are people who live in Korea for more than 20 years and can read and write in English. Five Korean people participated in the evaluation including one female and four males. Ages are between 24 and 35 where the average is 28.4. The participants have lived in Korea for from 20 years to 28 years and the average duration is 24.2.

Five American people participated in the evaluation as American human subjects including two females and three males. The ages are between 19 and 33, where the average is 25.8. All of them never lived outside of America except for short trips or travels.

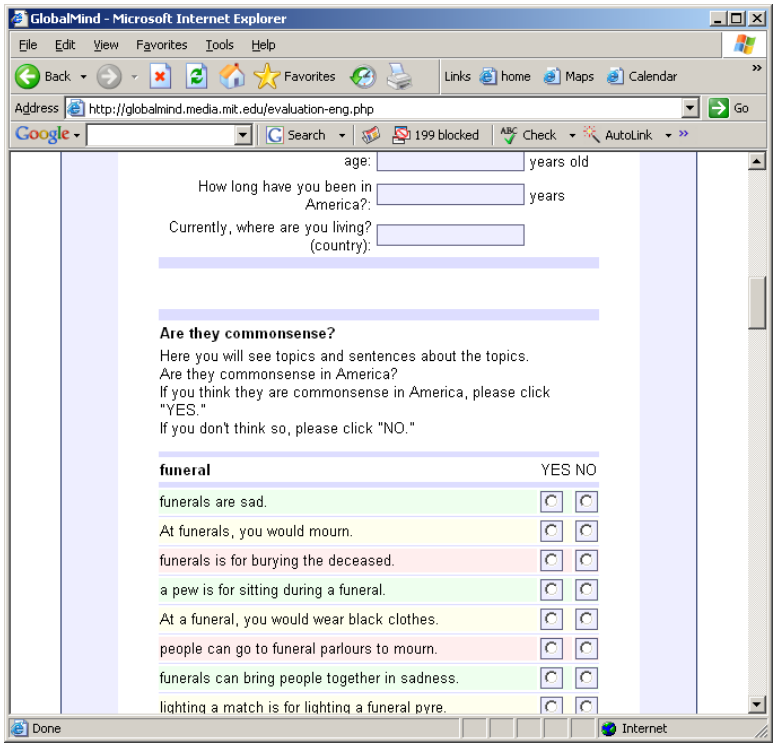
Survey Forms

The human decisions are done by on-line survey forms. Participants are asked to fill the survey forms out on-line. Figure 5-3 shows the screen shots of the on-line survey forms.

The survey forms show the sentences and check boxes with "yes" or "no." If participants think the sentence is common sense in their own culture, they mark "yes," and if the sentence if not commonsense, they mark "no."



(a) Korean form



(b) English form

Figure 5-3: DIM evaluation survey forms

5.2.2 Result

The whole answers of human participants can be found at Appendix A; Table A.5 shows the human decisions on the remaining set, and Table A.6 shows the human decisions on the subtracted set.

Among five participants in each group of Korean group and American group, if more than 60% participants agreed a sentence is "yes" then the sentences is regarded as "yes," and if more than 60% participants agreed a sentence is "no" then the sentence is regarded as "no." For a sentence, if American people answered "yes" but Korean people answered "no," the sentence is marked as "differences by human", and if both people answered "yes" then the sentence is marked as "similarities by human". The sentences that are judged as "no" by more than 60% of American participants are disregarded in this discussion because the basic assumption of this evaluation is all the English sentences are common sense for American people.

Table 5.4 shows the summary of human decisions on each set.

Table 5.4: Human decisions on each set

	the initial set	the remaining set	the subtracted set
Total	57	37	20
Number of "differences"	6	5	1
Number of "similarities"	51	32	19
Rate of "differences"	10.53%	13.50%	5.00%
Rate of "similarities"	89.47%	86.50%	95.00%

As discussed in Section 5.2.1, if the inference module functions as intended, the rate of "differences" of the initial set will be higher than that of the subtracted set and lower than that of the remaining set. The results show close resemblance of what is expected. The rate of "differences" in the initial set is 10.53%, which is higher than that of the subtracted set, 5.00%, and lower than that of the remaining set, 13.50%.

With the null hypothesis that the rate of the remaining set would be the lower than that of

the initial set, the p-value is 0.174. With the null hypothesis that the rate of the subtracted set would be higher than that of the initial set, the p-value is 0.227. Both p-values are not so strong, although both are less than 0.5. This result weakly supports that DIM works in the right direction it was intended.

In the best case, the rate of "differences" of the remaining set is close to 100% and the rate of "similarities" is close to 0%. However, the large rate, 86.50%, of the false positive in the remaining set does not indicates the failure of the inference module because they can be subtracted later when more data are accumulated in the databases. However, the false negative in the subtracted set is important, because once a sentence is mistakenly subtracted, it never returns to the remaining set. This result shows the very low rate of false negative in the subtracted set, 5.00%, which is cheerful. However, the fact that it is not 0% implies there is still room to improve the inference module.

Chapter 6

Applications

GlobalMind has two different kinds of inference modules. For applications, I developed two applications, one for each inference module: with Similarity Inference Module, Intercultural Dictionay has been developed; with Differences Inference Module, Personal Intercultural Assistant has been developed. Here I describe how the applications use the GlobalMind inference modules and how they are designed and implemented.

6.1 Intercultural Dictionary

Intercultural Dictionary is a dictionary to look up the most similar concept between two languages/cultures. The dictionary works between any two languages/cultures in GlobalMind and shows the similarity concepts of a given word between the two languages. For examples, if a user tries to look up the most similarity concept of English/America "fork" in Korea, the dictionary shows "JEOT GA RAK(chopsticks)" because "JEOT GA RAK(chopsticks)" in Korean/Korea is similar to "fork" in English/America in the point of that both of them are the most common utensil used for eating solid food and made of metal.

Intercultural Dictionary provides the main user interface for GlobalMind Similarity Inference Module. The user interface is written in Java Applet and requires three inputs: the

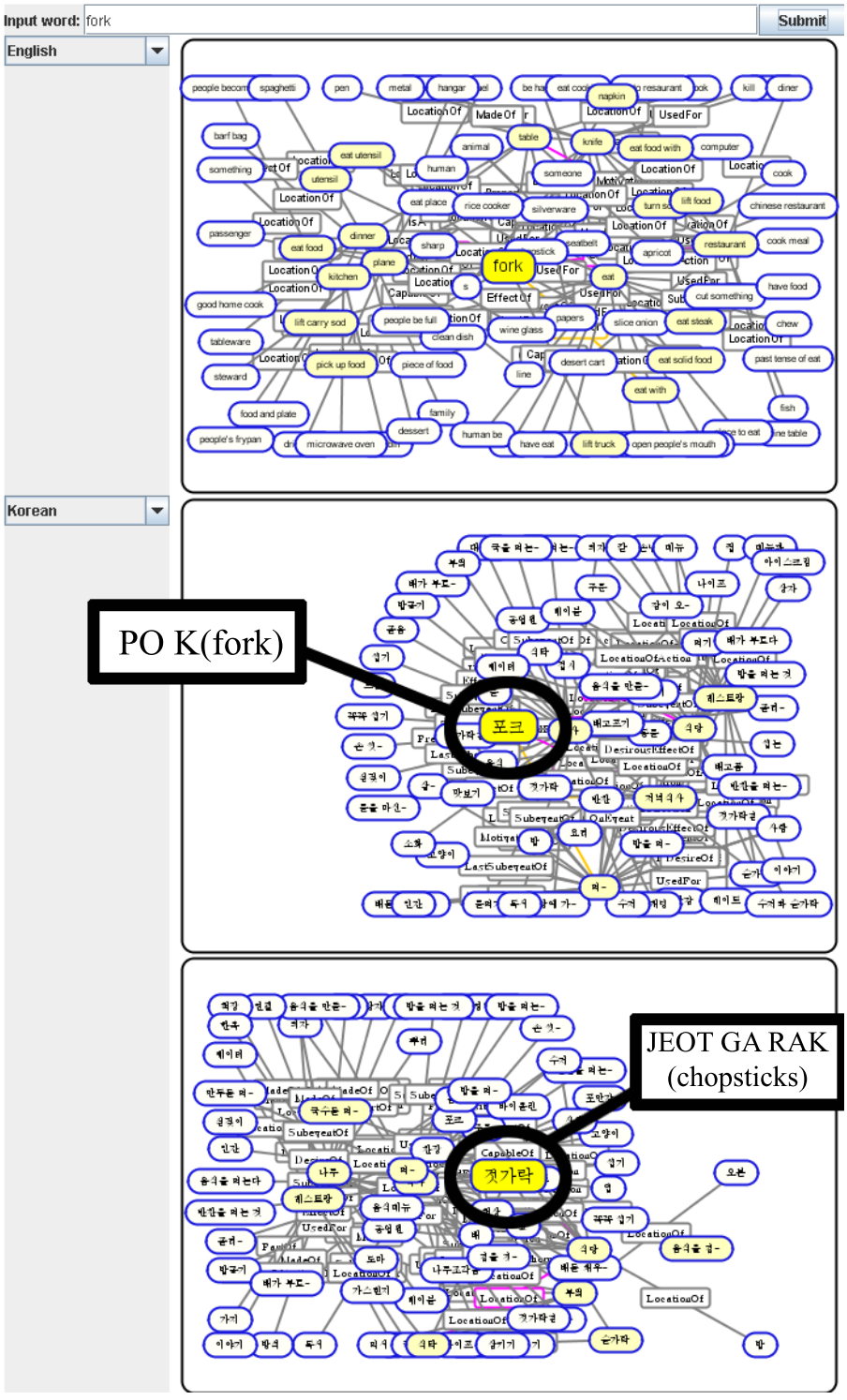


Figure 6-1: GlobalMind Intercultural Dictionary shows the result of “fork”

given language/culture, the target language/culture, and the give concept. For now, because of the limit of amount of accumulated common-sense knowledge, we assume that one language is involved in one culture; English common-sense knowledge is regarded as American common-sense, and Korean(language) common-sense knowledge is regarded as Korean(culture) common-sense knowledge. The output of the inputs is the list of the candidates of the most similar concepts in the target language/culture and the related networks used for the inference.

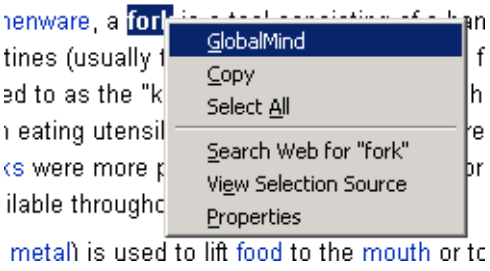
Figure 6-1 shows the Intercultural Dictionary serviced in the GlobalMind website. It shows "PO K(fork)" as the first candidate for the most similar concepts of "fork" and "JEOT GA RAK(chosticks)" as the second candidate. It also shows the contexts of each word.

This Intercultural Dictionary is also serviced as a FireFox extension. While surfing the Internet with a FireFox web browser, if a user highlights a phrase and clicks a right button of a mouse, then a user can see a GlobalMind button in the FireFox menu. (Figure 6-2(a)) If a user selects "GlobalMind," then GlobalMind Intercultural Dictionary pop up windows appears with the inference result of the highlighted phrase.

6.2 Personal Intercultural Assistant

Another application for GlobalMind is Personal Intercultural Assistant (PIA). PIA is intended to help foreign visitors to adapt themselves easier to the visiting countries. PIA helps users by providing the information of cultural differences between the users' original countries and the visiting countries.

PIA is developed on a cellular phone, Motoroal i870, using J2ME. A user can hold the cellular phone with PIA anywhere he wants and get helps from PIA anytime he needs. When the user runs PIA on his cellular phone, PIA will ask him his current situation. After entering the current situation of his, PIA will prompt the cultural differences on the current situation between two countries. For examples, if the user enters "restaurant" as



(a) Extension menu

(b) Popped-up GlobalMind dictionary

Figure 6-2: GlobalMind FireFox Extension

the current situation, and PIA knows that he is from Korea and he is now in the USA, PIA will show the information such as "please be aware of leaving tips for your waiter" which is not common-sense in Korea but common-sense in the USA.

Because a cellular phone can provide contexts information of users such as time, location, and contact lists, we can think about improving PIA to more context-based services. For examples, if the cellular phone is located at a restaurant at evening, PIA can assume that the user enters to a restaurant to have a dinner, and then provide better services before the user asks.

Figure 6-3: Personal Intercultural Assistant

Chapter 7

Future Work and Conclusion

7.1 Future work

In this section, I discuss several directions which can be considered in future work.

7.1.1 Improving Data Acquisition

The capability of GlobalMind mostly depends on the quantity and the quality of common-sense knowledge data in the GlobalMind database. Thus, to improve GlobalMind, it is necessary to accumulate more common-sense knowledge data.

GlobalMind has been using a website to gather the common-sense data and depending on endeavor of the Internet volunteers. However, the current voluntary participation has limitation; it can only collect the common-sense which can be recognized and can be written as a form of sentences, the process of typing in the common-sense sentences are reported as boring and even painful, and the accumulation is dependant on the factor we cannot control - the number of volunteers.

Thus, we can consider to find new ways to gather multilingual/multicultural common-sense

knowledge to improve GlobalMind.

To make it independent to the volunteers and to gather more common-sense knowledge in less time and efforts, Eslick [10] showed a web-mining robot collecting common-sense knowledge through the Internet could be a good solution. Eslick's web robot was built for English common-sense assertions for the original OpenMind database. However, we can consider the similar web robot to collect multilingual/multicultural common-sense.

Another way to improve the method of collecting common-sense is encouraging volunteers by making the process interesting and amusing. Ahn [3] showed the possibility of using games to train artificial intelligences. Similarly, we can consider amusing games to gather multilingual/multicultural common-sense knowledge data.

In addition, we can consider gathering other types of common-sense knowledge than sentences. Some knowledge which are too metaphysical or too physical cannot be recognized or written as a sentence. Thus, finding and designing new types of multilingual/multicultural common-sense representations other than sentences and predicates, and new methods to gather the new types of common-sense is an interesting future work.

7.1.2 Improving Machine Translation

As described in Section 1.1.2, language differences between two countries cannot be fully solved without consideration of cultural differences between them. Although much research has been done on machine translation at many institutions, the machine translation considering cultural differences has not been a popular research topic yet. Thus, one of interesting directions of further research is using GlobalMind to enhance traditional machine translators by adding the factor of cultural differences. For this purposes, GlobalMind provides several tools which can be used in machine translation: multilingual corpora and Similarity Inference Module to choose the similar concept between two cultures.

Another point of GlobalMind in machine translation field is that the approach used in GlobalMind is novel in the field of machine translation. While traditional machine translators depend on word-to-word mapping and limited scale bilingual corpora without consideration of cultural differences, GlobalMind approaches the same problems with link-to-link mapping, endlessly updated multilingual corpora, and context-based matching system with consideration of cultural differences. Adding this new point of view and new solution to classical solutions may result enhancement of the traditional machine translations.

7.1.3 User Interfaces and Applications

Although GlobalMind provides the fundamentals and tools for analyzing cultural similarities and differences, braiding the methods into human life requires more work. Thus, developing and designing appropriate and non-obtrusive user interfaces to assist people with the cultural differences when and where it is needed is another interesting future work.

Personal Intercultural Assistant shows one example of how to use GlobalMind for practical applications. It can help foreign visitors not to make rude mistakes caused from cultural differences. Cultural differences can also improve e-mail communication between two people in two different countries [11].

However, these applications depend on the small windows for typing while interactions between different countries are spread all over the environment. Thus, I suggest the concept of ubiquitous computing be combined with GlobalMind to improve people's interactions in the all directions.

7.2 Conclusion

The communication without misunderstanding is important in human interactions. However, it is difficult to avoid misunderstanding in the interactions between different countries

because people from different countries stand on different cultures and behave and analyze the other's behaviors based on different contexts.

This thesis described how large-scale common-sense knowledge databases and its inference modules can enrich the communication and the interactions among different countries. Although this research does not reach to the full implementation of the practical applications, it shows the potential of automated mechanisms to analyze the cultural differences and similarities through the GlobalMind project, a multilingual/multicultural common-sense knowledge database system, and provides the basic steps toward the further research on the enriched inter-cultural communication.

The quality of GlobalMind depends on the quantity of common-sense knowledge data. Thus, it may take a few more years for GlobalMind to gather enough data to make accurate analysis of cultures. On the other hand, there may come a new approach to these different cultures problems. However, the important thing is that these cultural differences problems should be approached and solved to enrich the interactions and to improve the quality of communication. And I believe this research contributes to improving communication among different countries by bringing the problems of different cultures to the center of communication problems and providing the foundations for the solutions.

Appendix A

Evaluation Data

A.1 Tables

Table A.1: SIM : Human decisions on the unconfirmed pairs 1

E word	K word	same	similar	related	no
First Candidate					
boy	SA RAM	0	2	5	0
children	YU CHI WON	1	0	4	1
city	DAE DO SI	2	4	1	0
example	YE MOON	4	2	1	0
Indian	A SI A	0	0	5	2
line	JWA SEOK	0	1	4	2
picture	PYO JI	2	0	2	3
place	JI GOO WI	0	0	2	5
point	CHOI JONG JEOM	0	2	4	1
side	CHEON JANG	0	0	1	6
story	HAK SEUB	1	0	0	6
word	MYUNG SA	0	3	4	0
1st Total		10	14	33	26
rate		12.05%	16.87%	39.76%	31.33%

Table A.2: SIM : Human decisions on the unconfirmed pairs 2

Second Candidate					
air	JI GOO WI	0	0	4	3
car	JA DONG CHA AN	0	5	2	0
city	HAN KOOK UI DO SI	0	0	6	1
earth	JI GOO WI	2	5	0	0
eye	EOL GOOL	0	0	6	1
face	MEO RI	0	1	6	0
family	A PA T	0	0	4	3
father	JEON DEUNG	0	0	0	7
hand	SON GA RAK	0	1	6	0
head	BAL	0	0	2	5
help	JIB EUL JIT DA	0	0	0	7
home	SIK TAK	0	0	4	3
Indian	S RI RAK KA	0	0	4	3
land	A SI A	0	0	4	3
mountain	DA RAM JUI	0	0	6	1
night	BAM E	5	2	0	0
number	CEOM PYU TEO	0	0	3	4
picture	DUIT JANG CEO VEO	0	0	3	4
place	DONG GUL AN	0	0	4	3
plant	SIM MUL EUN	3	4	0	0
river	GONG WON	0	0	5	2
school	HAK WON	1	2	2	2
sea	BAE	3	1	3	0
sentence	MAL	0	1	6	0
side	CHANG MOON	0	0	2	5
song	YEONG HWA BO GI	0	0	0	7
sound	RAK G ROOB	0	0	6	1
story	BAE UM	0	0	0	7
study	HAK GYO GA DA	0	1	5	1
time	BI HAENG GI	0	0	0	7
tree	PEOL P	0	0	6	1
word	MU EON GA	0	0	1	6
2nd Total		16	23	99	86
rate		7.14%	10.27%	44.20%	38.39%
Total		26	37	132	112
rate		8.47%	12.05%	43.00%	36.48%

Table A.3: DIM : the remaining set

	ID	sentence
Differences	011	funerals are sad.
	012	At funerals, you would mourn.
	013	funerals is for burying the deceased.
	014	a pew is for sitting during a funeral.
	015	At a funeral, you would wear black clothes.
	016	people can go to funeral parlours to mourn.
	017	funerals can bring people together in sadness.
	018	lighting a match is for lighting a funeral pyre.
	019	people can often wear black clothing to funerals.
	020	people can go to the funerals of people they knew.
	097	a restaurant is for eating.
	098	restaurants can serve food.
	099	a restaurant can serve wine.
	100	restaurants can employ waiters.
	101	a restaurant is for socializing.
	102	At the resturant, choose a menu.
	103	fancy restaurants are expensive.
	104	At the resturant, tip the waiter.
	105	a chef can may work at restaurant.
	106	a deli restaurant can serves food.
	107	restaurants can often serve salad.
	108	some restaurants can serve chicken.
	109	At the resturant, wait for a table.
	110	a restaurant bill can must be paid.
	111	restaurants can serve fine cuisine.
	112	a restaurant patron can order food.
	113	a restaurant table is for eating at.
	114	people can often eat in restaurants.
	115	a restaurant table is for dining at.
	137	waiters can serve a meal.
	138	a waiter can wait tables.
	139	waiter is part of wait staff.
	140	meat is for waiters to serve.
	141	waiting tables is for waiters.
	142	restaurants can employ waiters.
	143	waiter wants people to tip well.
	144	At the resturant, tip the waiter.
	145	waiters can serve food in a restaurant.
	146	waiters can serve meals in a restaurant.

Table A.4: DIM : the subtracted set

	ID	sentence
Similarities	021	a funeral home can prepares a deceased body for burial.
	022	dressing nice is for people who are going to a funeral.
	023	people can hugging near a cross. because i see a cross i presume this is a funeral.
	116	jews can eat in a kosher restaurant.
	117	busses can take you to a restaurant.
	118	restaurants can sell prepared meals.
	119	a restaurant table is for sitting at.
	120	At the resturant, ask for more water.
	121	turner can wanted his own restaurant.
	122	many americans can eat at restaurants.
	123	going to a restaurant is for the rich.
	124	a restaurant can often contains a bar.
	125	people can eat together in restaurants.
	126	waiters can serve food in a restaurant.
	147	At the resturant, you would call the waiter.
	148	waiters can spend a lot of time on their feet.
	149	a waiter can serves dinner to paying customers.
	150	setting a cup on a table is for a waiter to do.
	151	a waiter can serves food to people in a restaurant.
	152	using a calculator is for calculating waiters' tips.
	153	a waiter can serving a bottle of wine to some guests.
	158	If you want to order food then you should look for a waiter.
	161	If you want to serve customers then you should take a job as a waiter.
	162	a waiter can gives you your restaurant bill for eating at the restaurant.

Table A.5: DIM : Human decisions on the remaining set

Q	K1	K2	K3	K4	K5	A1	A2	A3	A4	A5	R
11	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
12	N	N	N	N	Y	Y	Y	Y	Y	Y	20
13	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
14	Y	N	Y	N	N	Y	Y	Y	Y	Y	40
15	N	Y	N	Y	Y	Y	Y	Y	Y	Y	60
16	Y	Y	Y	N	Y	Y	Y	Y	N	Y	80
17	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
18	N	N	Y	N	N	Y	N	Y	N	N	D
19	N	Y	N	Y	Y	Y	Y	Y	Y	Y	60
20	N	Y	N	Y	Y	Y	Y	Y	Y	Y	60
97	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
98	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
99	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
100	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
101	Y	N	Y	Y	Y	Y	Y	Y	Y	Y	80
102	Y	Y	Y	Y	Y	N	N	N	N	Y	D
103	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
104	N	N	Y	N	N	Y	Y	Y	Y	Y	20
105	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
106	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
107	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	80
108	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
109	Y	N	Y	N	Y	Y	Y	Y	Y	Y	60
110	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
111	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	80
112	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
113	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
114	Y	Y	N	Y	Y	Y	Y	Y	Y	Y	80
115	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
137	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
138	N	Y	N	Y	Y	Y	Y	Y	Y	Y	60
139	N	Y	N	Y	Y	Y	Y	Y	Y	Y	60
140	N	N	Y	Y	Y	Y	Y	Y	N	Y	60
141	Y	N	N	Y	Y	Y	Y	Y	Y	Y	60
142	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	80
143	N	N	Y	N	Y	Y	Y	Y	Y	Y	40
144	N	N	Y	N	N	Y	Y	Y	Y	Y	20
145	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
146	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	80
total											37
num of n(40)/y											5
per of n(40)/y											0.135
num of n(30)/y											3
per of n(30)/y											0.081
per of v/num											0.778

Table A.6: DIM : Human decisions on the subtracted set

Q	K1	K2	K3	K4	K5	A1	A2	A3	A4	A5	R
21	-	Y	-	Y	Y	Y	Y	Y	Y	Y	100
22	-	N	-	N	N	Y	Y	Y	Y	Y	0
23	-	N	-	N	N	N	N	Y	N	N	D
116	N	Y	-	Y	Y	Y	Y	Y	Y	Y	75
117	Y	Y	-	Y	Y	Y	N	Y	Y	Y	100
118	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	100
119	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	100
120	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	100
121	N	Y	-	N	Y	N	N	N	N	N	D
122	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	100
123	N	N	-	Y	N	N	N	Y	N	N	D
124	Y	N	-	Y	Y	Y	Y	Y	Y	Y	75
125	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	100
126	Y	Y	-	Y	Y	Y	Y	Y	Y	Y	100
147	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
148	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
149	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
150	-	N	N	Y	Y	Y	Y	Y	Y	Y	50
151	-	Y	Y	Y	Y	Y	Y	Y	Y	Y	100
152	-	N	N	N	N	Y	N	Y	N	Y	D
153	-	Y	N	Y	Y	Y	Y	Y	Y	Y	75
158	-	N	Y	Y	Y	Y	Y	Y	Y	Y	75
161	-	N	Y	N	Y	Y	N	Y	-	Y	50
162	-	Y	N	Y	Y	Y	Y	Y	Y	Y	75
total											20
num of n(40)/y											1
per of n(40)/y											0.050
num of n(30)/y											1
per of n(30)/y											0.050
per of v/num											0.838

Appendix B

Database Design

B.1 SQL commands for Database Creation

B.1.1 Group Tables

CREATE TABLE ‘groupmember‘ (
‘GMUID‘ int(10) unsigned NOT NULL auto_increment,
‘MUID‘ int(10) unsigned NOT NULL default ’0’,
‘GUID‘ int(10) unsigned NOT NULL default ’0’,
‘level‘ smallint(6) NOT NULL default ’0’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘GMUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘groups‘ (
‘GUID‘ int(10) unsigned NOT NULL auto_increment,
‘gname‘ varchar(30) NOT NULL default ”,
‘gtitle‘ varchar(255) NOT NULL default ”,
‘gtext‘ varchar(255) default NULL,

‘gowner‘ int(10) unsigned NOT NULL default ’0’,
‘gclose‘ smallint(6) NOT NULL default ’0’,
‘gpassword‘ varchar(41) default NULL,
‘gpasshint‘ varchar(255) default NULL,
‘gprefLanguage‘ varchar(3) NOT NULL default ’eng’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘GUID‘),
UNIQUE KEY ‘gname‘ (‘gname‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘rawdatagroup‘ (
‘RGUID‘ int(10) unsigned NOT NULL auto_increment,
‘RUID‘ int(10) unsigned NOT NULL default ’0’,
‘GUID‘ int(10) unsigned NOT NULL default ’0’,
‘valid‘ smallint(6) NOT NULL default ’0’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘RGUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

B.1.2 Administration Tables

CREATE TABLE ‘languageadmin‘ (
‘MUID‘ int(10) unsigned NOT NULL default ’0’,
‘lcode‘ char(3) NOT NULL default ”,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

B.1.3 Member Tables

```
CREATE TABLE ‘languagemember‘ (
‘LMUID‘ int(10) unsigned NOT NULL auto_increment,
‘MUID‘ int(10) unsigned NOT NULL default ’0’,
‘LUID‘ int(10) unsigned default NULL,
‘lcode‘ char(3) NOT NULL default ’’,
‘fluency‘ smallint(6) NOT NULL default ’0’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘LMUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘emails‘ (
‘email‘ varchar(50) default ’’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘members‘ (
‘MUID‘ int(10) unsigned NOT NULL auto_increment,
‘name‘ varchar(30) NOT NULL default ’’,
‘password‘ varchar(41) NOT NULL default ’’,
‘firstname‘ varchar(255) NOT NULL default ’’,
‘lastname‘ varchar(255) NOT NULL default ’’,
‘gender‘ char(1) NOT NULL default ’’,
‘prefLanguage‘ varchar(3) NOT NULL default ’eng’,
‘ccode‘ varchar(2) NOT NULL default ’’,
‘birthyear‘ smallint(6) NOT NULL default ’0’,
‘email‘ varchar(50) default ’’,
‘homepage‘ varchar(255) NOT NULL default ’’,
‘text‘ varchar(255) default ’’,
```

‘level‘ smallint(6) default ’0’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘MUID‘),
UNIQUE KEY ‘name‘ (‘name‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

B.1.4 Information Tables

CREATE TABLE ‘countries‘ (
‘ccode‘ varchar(2) NOT NULL default ”,
‘cname‘ varchar(255) NOT NULL default ”,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
UNIQUE KEY ‘ccode‘ (‘ccode‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘languages‘ (
‘LUID‘ int(10) unsigned NOT NULL auto_increment,
‘name‘ varchar(30) default ”,
‘englishname‘ varchar(30) default ”,
‘lcode‘ varchar(3) NOT NULL default ”,
‘used‘ smallint(5) unsigned default ’0’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘LUID‘),
UNIQUE KEY ‘lcode‘ (‘lcode‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘patternexamples‘ (
‘EUID‘ int(10) unsigned NOT NULL auto_increment,
‘TUID‘ int(10) unsigned NOT NULL default ’0’,

‘example‘ varchar(255) NOT NULL default ”,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘EUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

B.1.5 Data Tables

CREATE TABLE ‘nodes‘ (
‘NUID‘ int(10) unsigned NOT NULL auto_increment,
‘text‘ varchar(255) NOT NULL default ”,
‘lcode‘ varchar(3) NOT NULL default ”,
‘type1‘ varchar(20) default ”,
‘type2‘ varchar(20) default ”,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
‘valid‘ smallint(5) unsigned default ’1’,
PRIMARY KEY (‘NUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘patterns‘ (
‘RUID‘ int(10) unsigned NOT NULL default ’0’,
‘TUID‘ int(10) unsigned default NULL,
‘function‘ varchar(30) NOT NULL default ”,
‘node1‘ varchar(255) NOT NULL default ”,
‘node2‘ varchar(255) default NULL,
‘NUID1‘ int(10) unsigned default NULL,
‘NUID2‘ int(10) unsigned default NULL,
‘valid‘ smallint(6) NOT NULL default ’0’,
‘author‘ int(10) unsigned default NULL,
‘lcode‘ varchar(3) NOT NULL default ”,

‘ccode‘ varchar(2) NOT NULL default ”,
‘translated‘ smallint(6) NOT NULL default ’0’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘patterntemplates‘ (
‘TUID‘ int(10) unsigned NOT NULL auto_increment,
‘function‘ varchar(30) NOT NULL default ”,
‘lcode‘ varchar(3) NOT NULL default ”,
‘template‘ varchar(255) NOT NULL default ”,
‘node1type1‘ varchar(20) default ”,
‘node1type2‘ varchar(20) default ”,
‘node2type1‘ varchar(20) default ”,
‘node2type2‘ varchar(20) default ”,
‘used‘ smallint(5) unsigned default ’1’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
‘type‘ varchar(30) default ’single’,
PRIMARY KEY (‘TUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘rawdata‘ (
‘RUID‘ int(10) unsigned NOT NULL auto_increment,
‘text‘ varchar(255) NOT NULL default ”,
‘author‘ int(10) unsigned NOT NULL default ’0’,
‘valid‘ smallint(6) NOT NULL default ’0’,
‘activity‘ varchar(255) default ”,
‘related‘ int(10) unsigned default NULL,
‘lcode‘ varchar(3) NOT NULL default ”,
‘ccode‘ varchar(2) NOT NULL default ”,
‘translated‘ smallint(6) NOT NULL default ’0’,

‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘RUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

CREATE TABLE ‘rawdatalanguage‘ (
‘RLUID‘ int(10) unsigned NOT NULL auto_increment,
‘RUID1‘ int(10) unsigned NOT NULL default ’0’,
‘lcode1‘ char(3) NOT NULL default ”,
‘ccode1‘ char(2) NOT NULL default ”,
‘RUID2‘ int(10) unsigned NOT NULL default ’0’,
‘lcode2‘ char(3) NOT NULL default ”,
‘ccode2‘ char(2) NOT NULL default ”,
‘author‘ int(10) unsigned NOT NULL default ’0’,
‘date‘ timestamp NOT NULL default CURRENT_TIMESTAMP on update CURRENT_TIMESTAMP,
PRIMARY KEY (‘RLUID‘)
) ENGINE=InnoDB DEFAULT CHARSET=utf8;

Appendix C

Reader Biography

C.1 Professor Sung Hyon Myaeng

Prof. Sung Hyon Myaeng is currently a professor at Information and Communications University (ICU), Korea. Prior to this appointment, he was a faculty at Chungnam National University, Korea, and Syracuse University, USA, where he was granted tenure. He has served on program committees of many international conferences in the areas of information retrieval, natural language processing, and digital libraries, including his role as a chair for ACM SIGIR, 2002, and for AIRS, 2004. He is an editorial board member for Information Processing and Management, Journal of Natural Language Processing, and Journal of Computer Processing of Oriental Languages for which he is the information retrieval (IR) area chair. He was an associate editor for ACM Transactions on Asian Information Processing from 1999 to 2003. He is currently the chair of SIG-HLT (Human Language Technology), Korea Information Science Society. He has published numerous technical articles on elicitation of semantic relations from text, conceptual graph-based IR, cross-language IR, automatic summarization, text categorization, topic detection and tracking, and distributed IR in the context of digital libraries. Recently he has embarked on a project to use commonsense knowledge base for IR, text mining, and e-health applications, which is related to his involvement in semantic web research.

Bibliography

[1] Adler, N.J. and John L. Graham, "Cross-Cultural Interaction: the International Comparison Fallacy?," *Journal of International Business Studies,* **20**(3), (1989): 515-537.

[2] Anacleto, J., Lieberman, H., Tsutsumi, M., Neris, V., Carvalho, A., Espinosa, J., and Zem-Mascarehnhas, S., "Can Common Sense Uncover Cultural Differences in Computer Applications?," *World Computer Congress,* (2006).

[3] Ahn, L. and Dabbish, L., "Labeling images with a computer game," *Proc. of the SIGCHI conference on Human factors in computing systems,* (New York, NY: ACM Press, 2004): 319–326.

[4] Brown, P.F., Cocke, J., Pietra, S.D, Pietra, V.J.D., Jelinek, F., Lafferty, J.D., Mercer, R.L., and Roossin, P.S., "A Statistical Approach to Machine Translation," *Computational Linguistics,* **16**(2), (1990): 79–85.

[5] Brown, R.D., "Example-Based Machine Translation in the Pangloss System," *Proc. of Computational linguistics,* (1996): 169–174.

[6] ConceptNet, http://www.conceptnet.org/, (Retrieved November 09, 2005).

[7] Condon, J.C., "Perspective for the conference," *Intercultural Encounters with Japan,* J.C. Condon and M Saito, ed., (Tokyo: Simul Press, 1974).

[8] Dictionary.com, http://www.m-w.com/cgi-bin/dictionary, (Retrieved November 09, 2005).

[9] ESL Desk - Learn English as a Second Language, http://www.esldesk.com/esl-quizzes/most-used-english-words/words.htm, (Retrieved August 1, 2006).

[10] Eslick, I., http://web.media.mit.edu/ eslick/, (Retrieved June 26, 2006).

[11] Espinosa, J., http://agents.media.mit.edu/projects/culture/, (Retrieved November 09, 2005).

[12] Ethnologue Languages of the World, http://www.ethnologue.com/language_index.asp, (Retrieved June 29, 2006).

[13] Euzenat, J., "an API for Ontology Alignment," *Proc. of the International Semantic Web Conference,* (2004): 698–712.

[14] Fellbaum, F. C., "WordNet an electronic Lexical Database," (Cambridge, MA: The MIT Press, 1998).

[15] Gentner, D., "Structure-mapping: A theoretical framework for analogy," *Cognitive Science* **7**(2), (1983): 155-170.

[16] GlobalMind Website, http://globalmind.media.mit.edu, (Retrieved June 26, 2006).

[17] Google Translate, http://translate.google.com/translate_t, (Retrieved November 09, 2005).

[18] Herring, R.D., "Nonverbal Communication: A Necessary Component of Cross-Cultural Counseling," *Journal of Multicultural Counseling and Development,* **18**(4), (1990): 172-179.

[19] Hofstadter, D. R., "Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought," (Basic Books, 1996).

[20] Hofstadter, D. R., "Le Ton Beau De Marot: In Praise of the Music of Language," (Basic Books, 1999).

[21] Jurafsky, D. and Martin, H., "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition," (Cambridge, MA: The MIT Press, 2000).

[22] Khaslavsky, J., "Integrating culture into interface design," *CHI 98 conference summary on Human factors in computing systems,* (1998): 365–366.

[23] Knight, K. and Luk, S.K., "Building a Large-Scale Knowledge Base for Machine Translation," *American Association for Artificial Intelligence,* (1994): 773–778.

[24] Lakoff, G., "Women, Fire and Dangerous Things," (University of Chicago Press, 1987).

[25] Lenat, D. B., "The dimensions of Context space," (Austin, TX: Cycorp, 1998).

[26] Liu, H. and Singh, P., "ConceptNet: a Practical Commonsense Reasoning Toolkit" *BT Technology Journal,* **22**(4), (Kluwer Academic Publishers, 2004): 211–226

[27] Manning, D. and Schtoze, H., "Foundations of Statistical Natural Language Processing Christopher," (Cambridge, MA: The MIT Press, 1999).

[28] Marcus, A. and Gould, E. W, "Crosscurrents: Cultural Dimensions and Global Web-User Interface Design," *ACM Interactions Association for Computer Machinery Inc.,* **7**(4) (2000):32–46.

[29] Merriam-Webster Online Dictionary, http://www.m-w.com/cgi-bin/dictionary, (Retrieved November 09, 2005).

[30] MontyLingua, http://web.media.mit.edu/ hugo/montylingua/, (Retrieved November 09, 2005).

[31] Mueller, E. T., "Natural language processing with ThoughtTreasure," (New York: Signiform, 1998).

[32] Munter, M., "CROSS-CULTURAL COMMUNICATION FOR MANAGERS," *Business Horizons,* **36**(3), (1993).

[33] Noy, N. F., "Semantic integration: a survey of ontology-based approaches," *ACM SIGMOD Record,* **33**(4), (New York, NY: ACM Press, 2004):65–70.

[34] OpenMind Common Sense, http://commonsense.media.mit.edu/, (Retrieved November 09, 2005).

[35] Oxford English Dictionary, http://dictionary.oed.com/, (Retrieved November 09, 2005).

[36] Russo, P. and Boor, S., "How fluent is your interface?: designing for international users," *Proc. of the SIGCHI conference on Human factors in computing systems,* (1993): 342–347.

[37] Sawyer, J. and Guetzkow, H., "Bargaining and Negotiation in International Relations," *International Behavior: a Social-Psychological Analysis,* ed. Herbert C. Kelman, (New York: Holt, Rinehart and Winston, 1965): 464–520.

[38] Scheff, T.J., "Is Accurate Cross-Cultural Translation Possible?," *Current Anthropology,* **28**(3), (1987): 365.

[39] Sechrest, L., Fay, T.L., and Zaidi, S.M.H., "Problems of Translation in Cross-Cultural Research," *Journal of Cross-Cultural Psychology,* **3**(1), (1972): 41–56.

[40] Singh, P., "The Public Acquisition of Commonsense Knowledge," *Proc. of AAAI Spring Symposium on Acquiring (and Using) Linguistic (and World) Knowledge for Information Access,* (2002).

[41] Sumita, E. and Iida, H., "Experiments and prospects of Example-Based Machine Translation," *Proc. of Association for Computational Linguistics,* (1991): 185–192.

[42] Tomita, M. and Carbonell, J.G., "Another Stride Towards Knowledge-Based Machine Translation," *Proc. of Computational linguistics,* (1986): 663–668.

[43] Uren, E., Howard, R., and Perinotti, T., "Software Internationalization and Localization: An Introduction," (Van Nostrand Reinhold, 1993).

[44] Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venugopal, A., Zhao, B., and Waibel, A., "The CMU statistical machine translation system," *Proc. of MT Summit,* (2003): 402–409.

[45] Yahoo Korea English Dictionary, http://dic.yahoo.com/, (Retrieved July 28, 2006).