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

by

Hyemin Chung, Henry Lieberman, and Walter Bender

MIT Media Lab
20 Ames Street
Cambridge, MA 02139
{ence, lieber, walter} @ media.mit.edu

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 paper approaches the problems with automatic computation of cultural differences and similarities. GlobalMind provides commonsense databases of various countries and languages and two inference modules to analyze and compute the cultural differences and similarities from the databases. This paper describes the design of GlobalMind databases, the implementation of its inference modules, and the results of an evaluation of GlobalMind.

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.

Cultural Differences

In cross-cultural interactions, people should consider and understand the cultural background of each other in order to have successful interactions [1]. Expected behaviors, signals, and contexts of communication differ by cultural backgrounds of communicators. 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 [26]. Herring [11] showed cross-cultural counselors should understand cultural differences and apply those differences to their non-verbal communication styles to avoid misunderstandings. Condon [4] 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.

Language Differences

The language differences have been researched and studied by many people from linguistic researchers to elementary school students. The efforts to solve the language difference problem with automated mechanisms have resulted in many different kinds of mechanisms of machine translation. While these mechanisms have solved many parts of the problem, there still remain problems, and many of these remaining problems cannot be solved without consideration of cultural differences.

There have been many discussions about if an accurate translation between two different cultures is possible [27]. 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 [28]. Munter [22] 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." The existence or absence of the word in languages is also closely related to the existence or absence of the concept itself in cultures. Although this problem is grounded in language differences, it cannot be solved without understanding each other's cultures.

Some expressions with the same meanings can be used totally differently between cultures, and the other expressions with different meanings can be used for the same uses in different 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.

Thus, it is necessary to consider the cultural differences as well as the language differences when translating languages.

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.

GlobalMind provides programming tools for analyzing cultural contexts to reduce the problems described above and to improve the quality of cross-cultural communication. GlobalMind consists of 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.

Automated Mechanisms for Cultural Contexts Analysis

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, for the first time, a possibility of assistant programs to improve the understanding of cultural differences. However, it has limitation in both depth and breadth of data because some of the steps to process the comparisons among the different databases and the topics to be compared were not automatized, not automatically computed. With manual input, even though in some of the steps, it is still difficult to see how to extend and generalize the work. The credit of this work is that the authors proved that it is plausible to do automatic comparisons among different cultures using OMCS knowledge base.

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.

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 work 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, we 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 [29], and more than 700,000 items as of November 2005 [24]. 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.

Context-Based Analysis

Understanding contexts is an important key for successful cross-cultural communication. To improve context analysis, GlobalMind uses context-based approach. Here, the term "context" is not 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 cultural differences can be represented with different contexts. For example, 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" and "noodle soup."

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.

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.

GlobalMind uses 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)."

2 Background and Related Work

The research in this paper 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 [17], and ThoughtTreasure [21]. Cyc has collected knowledge only in English, and with little thought to cultural differences. Cyc does have a mechanism for establishing contexts [17] 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.

2.1 Cultural Issues and Interfaces

Aaron Marcus [20] and others have written extensively on the need for cultural sensitivity in user-interface design. Many people including Russo and Boor [25], and Khaslavky [15] 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 [30]. 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 [14]. 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. 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. WordNet [7] 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 [19] as representations of context, to use in word sense disambiguation. 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 [16]. Metaphors can be considered as generalizations of commonsense situations, we show in this paper how these generalizations can be carried over from one language and culture to another.

2.3 Ontology Alignment

Ontology Alignment is an active area of research in Artificial Intelligence [6] [23]. 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.

2.4 Analogies

Finally, much work in Artificial Intelligence concerns analogies. The classic reference is [8]. 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 [13] presents a delightful tour of the importance of analogy and metaphor in language translation. In [12], he and his colleagues explore computational and statistical mechanisms of analogy.

3 Design and Implementation

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 [18] 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.

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.

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 [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 paper the size of database or the number of common-sense items 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.

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 countries, 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. Figure 1 shows a snapshot of a network of "Shampoo."

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 2 shows the concept of a network of networks, the final form of GlobalMind database.

In GlobalMind, there are two different kinds of bilingual connections. The first one is the established bilingual connections, which are manually entered by bilingual volunteers or collected from bilingual manuscripts such as dictionaries. The number of the established bilingual connections is relatively few compared to the number of GlobalMind common-sense knowledge data because they are manually entered. The other kind of connections is the inferred bilingual connections, which are automatically computed by GlobalMind inference modules. Based on the established connections, GlobalMind automatically computes the relationship between any node in one language/culture network and other nodes in another language/culture network. The inference method will be described below. In this paper, if we refer the bilingual connection, it usually means the established bilingual connection.

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.

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 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 shows the concept of the expand-and-contract method.

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 us define several terms here. When there is a given node, we 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.

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 established bilingual connections between two language/culture networks.

SIM finds the estabilshed bilingual connections between the given sub-network and the target network. Bilingual connections mean two predicates each of which is located in each network and both of which have similar meanings. The bilingual connections are entered manually and thus much fewer and sparser than monolingual nodes and connections.

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.

Contracting the Sub-network

After finding the matching sub-network, we now have a pair of sub-networks: a given subnetwork 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 us 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.

The first factor considered in the weight system 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 [18]. 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 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.

Another factor, which can be considered but not implemented in GlobalMind yet, is the kind of relationships. 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. 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 comparison starts from the bilingual connections. At first, SIM searches the bilingual connections between two subnetworks. 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 as weights.

At first, SIM finds bilingually connected nodes in sub-networks. From the nodes, SIM
compares 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.

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 in [2] there were the first attempts to approach the cultural difference problems using OMCS, 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 commonsense is removed; (4) after comparing and removing, remained sub-networks are cultural differences between two cultures about the given situation.

Sub-network Extraction

The situation is usually written only in one language. For convenience in writing, here we will assume that the given situation is written in English.

When the situation is nodes written in English, we can easily extract the American subnetwork 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 [10]. 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 paper, 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.

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 [10] 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 1 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 is at least one matched word in each node and the links are the same, the predicate A and B are the shared common sense.

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, we hope this problem will be resolved as the database is enlarged.

3.3 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 [29]. 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.

To gather multilingual/multicultural common-sense knowledge, we built and launched GlobalMind website based on OpenMind website. GlobalMind website [9] 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.

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 2 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.

4 Evaluation

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.

4.1 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 concept, 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.

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.”

Test Concept Sets

The given concepts were chosen from the 300 most frequently used English words [5]. 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.

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.”

Result

Table 3 shows the answers of subjects. Count means how many times each answer selected by human subjects. Because there are 61 pairs and seven human subjects, the maximum count of answers for the word pairs is 427.

As described above, “same” and “similar” 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.

The result shows that the 76.47% of word pairs have the same concepts, and the 9.18% pairs have the similar concepts. The “matched” pairs are 85.65% of the whole word pairs which is higher than 50%. The rate of “unmatched” word pairs is 14.35% which is lower than 50%. With the null hypothesis of that the rate of the “matched” word pairs will be lower than 50%, and the alternative hypothesis of that the rate of “matched” word pairs will be higher than that of “unmatched” word 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 word pairs 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” word 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.

The word pairs generated by GlobalMind SIM were also compared to the Yahoo English-Korean dictionary [31]. 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 candidate pairs out of the 61 word pairs are “confirmed” by the dictionary. “Unconfirmed” pairs include wrong inferences and indirect inferences.

Here our 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 4 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.

4.2 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 commonsense 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 commonsense 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 4 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 4(a) shows the initial knowledge set that are not processed yet. In ideal case, as subtracting the set temporarily looks like Figure 4(b) and finally looks like Figure 4(c). In the bad case, it can subtract not-shared common sense by mistakes and it will look like Figure 4(d).

DIM determines the shared knowledge by data accumulated in the databases; if knowledge is located in both of American and Korean databases, it is 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 4, this evaluation will measure to which DIM is closer between Figure 4(b) or Figure 4(d) rather than between Figure 4(b) and 4(c).

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.

DIM processed the initial set and made the remaining set with 37 sentences and the subtracted set with 20 sentences.

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. 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.”

Result

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.

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 in Table 5 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.

5 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 paper 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 we 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.

Acknowledgements

We'd like to thank the late Push Singh for his great ideas, Wonsik Kim for his contribution to programming, and Samsung Lee Kun Hee Scholarship Foundation for financial supports.

References

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

[4] Condon, J.C., “Perspective for the conference,” *Intercultural Encounters with Japan,* J.C. Condon and M Saito, ed., (Tokyo: Simul Press, 1974).
[5] ESL Desk - Learn English as a Second Language, http://www.esldesk.com/esl-quizzes/mostused-english-words/words.htm, (Retrieved August 1, 2006).
[6] Euzenat, J., “an API for Ontology Alignment,” *Proc. of the International Semantic Web Conference,* (2004): 698-712.
[7] Fellbaum, F. C., “WordNet an electronic Lexical Database,” (Cambridge, MA: The MIT Press, 1998).
[8] Gentner, D., “Structure-mapping: A theoretical framework for analogy,” *Cognitive Science* 7(2), (1983): 155-170.
[9] GlobalMind Website, http://globalmind.media.mit.edu, (Retrieved June 26, 2006).
[10] Google Translate, http://translate.google.com/translate t, (Retrieved November 09, 2005).
[11] Herring, R.D., “Nonverbal Communication: A Necessary Component of Cross-Cultural Counseling,” *Journal of Multicultural Counseling and Development,* 18(4), (1990): 172-179.
[12] Hofstadter, D. R., “Fluid Concepts & Creative Analogies: Computer Models of the Fundamental Mechanisms of Thought,” (Basic Books, 1996).
[13] Hofstadter, D. R., “Le Ton Beau De Marot: In Praise of the Music of Language,” (Basic Books, 1999).
[14] 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).
[15] Khaslavsky, J., “Integrating culture into interface design,” *CHI 98 conference summary on Human factors in computing systems,* (1998): 365-366.
[16] Lakoff, G., “Women, Fire and Dangerous Things,” (University of Chicago Press, 1987).
[17] Lenat, D. B., “The dimensions of Context space,” (Austin, TX: Cycorp, 1998).
[18] Liu, H. and Singh, P., “ConceptNet: a Practical Commonsense Reasoning Toolkit” *BT Technology Journal,* 22(4), (Kluwer Academic Publishers, 2004): 211-226
[19] Manning, D. and Schtoze, H., “Foundations of Statistical Natural Language Processing Christopher,” (Cambridge, MA: The MIT Press, 1999).
[20] Marcus, A. and Gould, E. W, “Crosscurrents: Cultural Dimensions and GlobalWeb-User Interface Design,” *ACM Interactions Association for Computer Machinery Inc.,* 7(4) (2000):32-46.
[21] Mueller, E. T., “Natural language processing with ThoughtTreasure,” (New York: Signiform, 1998).
[22] Munter, M., “CROSS-CULTURAL COMMUNICATION FOR MANAGERS,” *Business Horizons,* 36(3), (1993).
[23] Noy, N. F., “Semantic integration: a survey of ontology-based approaches,” *ACM SIGMOD Record,* 33(4), (New York, NY: ACM Press, 2004):65-70.
[24] OpenMind Common Sense, http://commonsense.media.mit.edu/, (Retrieved November 09, 2005).
[25] 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.
[26] Sawyer, J. and Guetzkow, H., “Bargaining and Negotiation in International Relations,”

InternationalBehavior: a Social-Psychological Analysis, ed. Herbert C. Kelman, (New York: Holt, Rinehart and Winston, 1965): 464-520.
[27] Scheff, T.J., "Is Accurate Cross-Cultural Translation Possible?," *Current Anthropology,* **28**(3), (1987): 365.
[28] 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.
[29] 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).
[30] Uren, E., Howard, R., and Perinotti, T., "Software Internationalization and Localization: An Introduction," (Van Nostrand Reinhold, 1993).
[31] Yahoo Korea English Dictionary, http://dic.yahoo.com/, (Retrieved July 28, 2006).

Table 1 Comparison of two predicates in different languages

<table>
<tr><td rowspan="3">Predicate A</td><td>original</td><td>결혼식 (GYUL HON SIK) -> OnEvent -> 웨딩드레스를 입다 (WE DING D RE S RUL IB DA)</td></tr>
<tr><td>original meaning</td><td>wedding -> OnEvent -> wear wedding dress</td></tr>
<tr><td>machine translated</td><td>**wedding** ceremony -> OnEvent -> the [wey] [ting] puts on the **dress**</td></tr>
<tr><td colspan="2">Predicate B</td><td>**wedding** -> OnEvent -> wear wedding **dress**</td></tr>
<tr><td colspan="2">Shared Word</td><td>**wedding** -> OnEvent -> **dress**</td></tr>
</table>

Table 2 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
	Etc	467
	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

Table 3 Human answers for SIM evaluation

Relationship		Count		Rate	
Same	Matched	325	364	76.47%	85.65%
Similar		39		9.18%	
Related	Unmatched	35	61	8.24%	14.35%
Not related		26		6.12%	
Total		425		100%	

Table 4 Human answers for the unconfirmed pairs

		Total Answers		Rate	
Matched	Same	10	24	12.05%	28.92%
	Similiar	14		16.87%	
Unmatched	Related	33	59	39.76%	71.08%
	Not at all	26		31.33%	
Total			63		100%

Table 5 Human decisions on each set

	Count			Rate		
	Similarities	Differences	Total	Similarities	Differences	Total
Initial Set	51	6	57	89.47%	10.53%	100%
Remaining Set	32	5	37	86.50%	13.50%	100%
Subtracted Set	19	1	20	95.00%	5.00%	100%

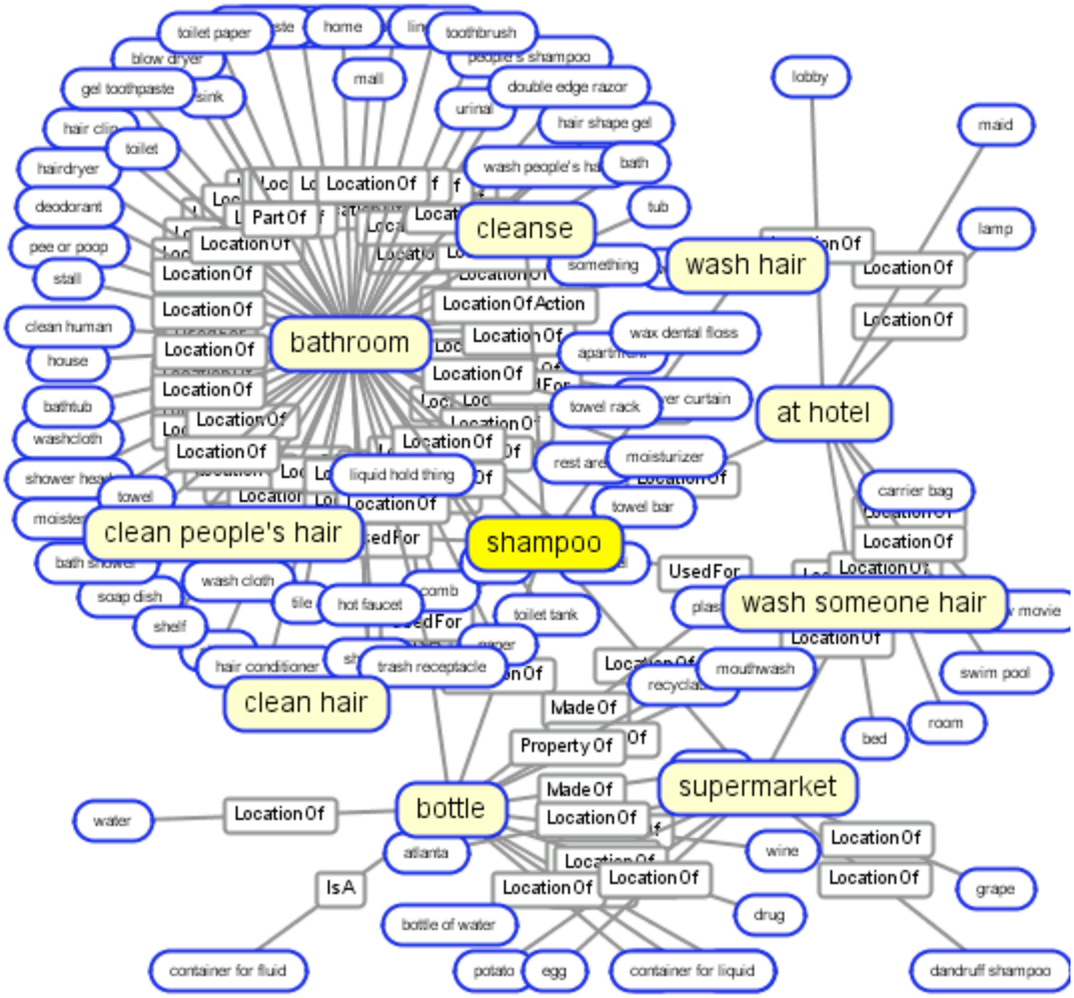

Figure 1 Snapshot of GlobalMind network about shampoo

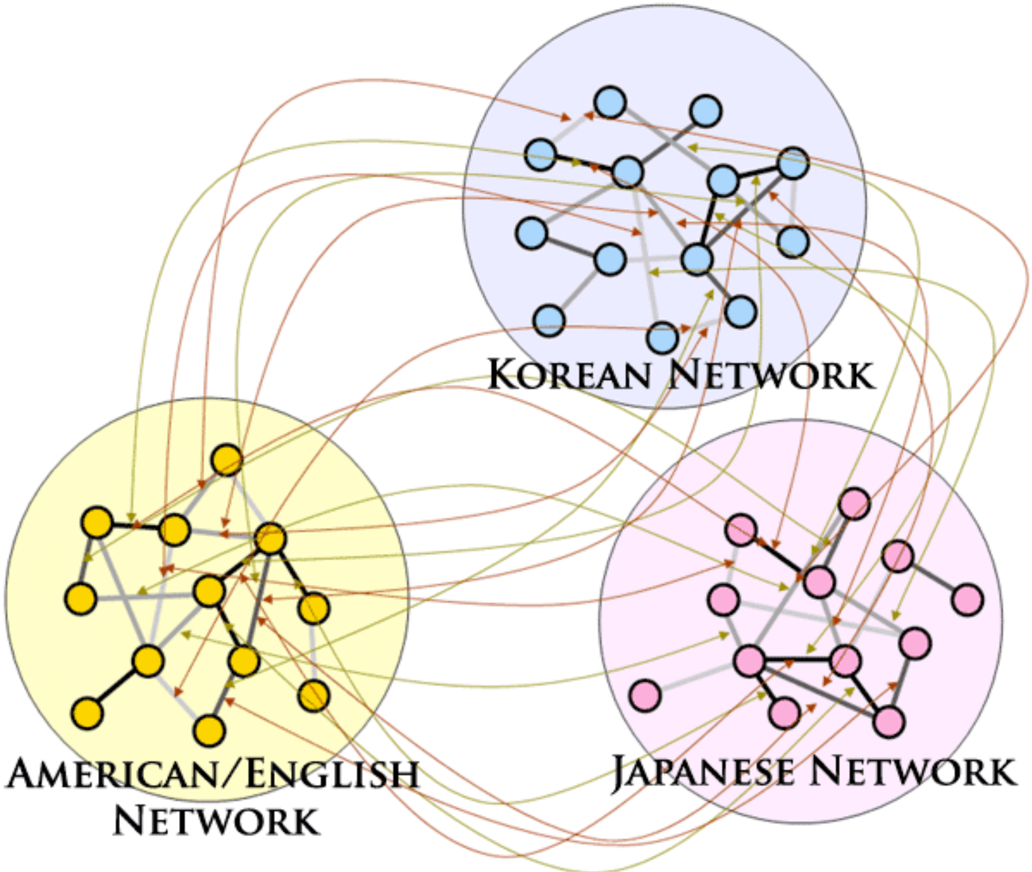

Figure 2 Conceptual image of GlobalMind global network

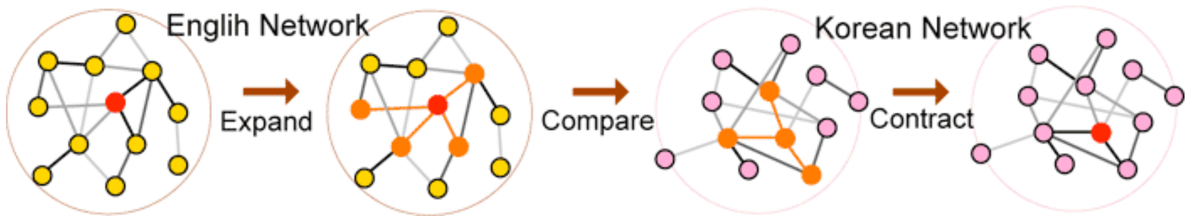

Figure 3 Expand and contract method

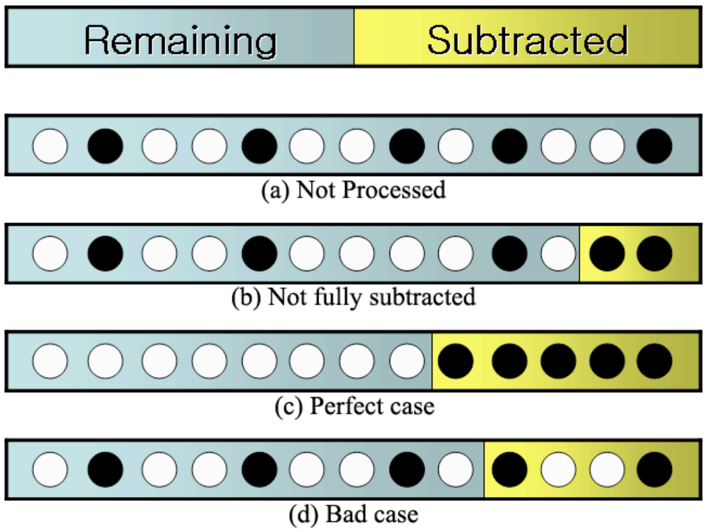

Figure 4 GlobalMind DIM processes and performance