IEICE TRANS. INF. & SYST., VOL.E91–D, NO.11 NOVEMBER 2008

PAPER Special Section on Knowledge, Information and Creativity Support System

Assisting Pictogram Selection with Categorized Semantics

Heeryon CHO[†a)], ***Nonmember***, **Toru ISHIDA**[†], ***Fellow***, **Satoshi OYAMA**[†], **Rieko INABA**[††], ***Members***, ***and*** **Toshiyuki TAKASAKI**[†,†††], ***Nonmember***

SUMMARY Since participants at both end of the communication channel must share common pictogram interpretation to communicate, the pictogram selection task must consider both participants' pictogram interpretations. Pictogram interpretation, however, can be ambiguous. To assist the selection of pictograms more likely to be interpreted as intended, we propose a *categorical semantic relevance measure* which calculates how relevant a pictogram is to a given interpretation in terms of a given category. The proposed measure defines similarity measurement and probability of interpretation words using pictogram interpretations and frequencies gathered from a web survey. Moreover, the proposed measure is applied to categorized pictogram interpretations to enhance pictogram retrieval performance. Five pictogram categories used for categorizing pictogram interpretations are defined based on the five first-level classifications defined in the Concept Dictionary of the EDR Electronic Dictionary. Retrieval performances among not-categorized interpretations, categorized interpretations, and categorized and weighted interpretations using semantic relevance measure were compared, and the categorized semantic relevance approaches showed more stable performances than the not-categorized approach.

key words: *semantic relevance, categorization, EDR, pictogram*

1. Introduction

Advances in information communication technology have enabled ordinary people to easily create, publish, and share various kinds of images such as photographs, movies, and illustrations, leading to a vast amount of image data to accumulate on the World Wide Web. Meanwhile, tag-based content management applications* have come into wide use, and user added tags, a prevalent form of metadata, are incorporated into image search process to assist image retrieval.

Among various image management applications that incorporate tags, we focus on a pictogram email system which allows children to communicate to each other using pictogram messages [1]. Existing pictogram communication systems such as Minspeak [2] and IconText [3] use a fixed set of icons and system-defined sentence generation procedures to create pictogram messages. By contrast, the pictogram email system [1] we deal with uses an open set of pictograms where new pictograms are continuously added to the existing set of pictograms. The email system provides a two-dimensional canvas interface where a user can freely place one or more pictograms onto the canvas to create pictogram messages; no system-defined pictogram sentence generation procedure is imposed on the user. While the pictogram sentence creation strategies of the existing systems [2], [3] utilize a fixed set of pictograms and predefined sentence generation procedures to generate clearly defined pictogram sentences, our system [1] uses an unfixed set of pictograms as candidates for conveying intended meaning, and so the selection of the most relevant pictogram becomes the sentence creation strategy. Therefore, we focus on the pictogram selection stage where children select individual pictograms to create pictogram messages.

Pictogram is an icon which has a clear pictorial similarity with some object [4], and one who can recognize the object depicted in the pictogram can interpret the meaning associated with the object. Pictorial symbols, however, are not universally interpretable. A simple design like an arrow is often used to show direction, but there is no reason to believe that arrows suggest directionality to all people; they might also be taken as a symbol for war or bad luck [5]. Since the selection of pictogram in the pictogram email system is done with the purpose of conveying certain meaning to the communicating counterpart, a pictogram that was selected must carry intended meaning to both the sender and receiver of communication; that is, selected pictogram must be relevant to participants at both end of communication channel in order for the pictogram communication to be successful.

To assist pictogram selection, we propose a *categorical semantic relevance measure*, which calculates how relevant a pictogram is to a given interpretation using categorized pictogram interpretations. Related researches that utilize tags (which could be viewed as a kind of interpretation given by the user) unifies browsing by tags and visual features for intuitive exploration of image databases [6] or helps users browse large scale annotations in semantic, hierarchical, and efficient way [7]. In [6], navigation within the image database is augmented by combining image tags with visual features of the images while [7] utilizes tags and

Manuscript received April 4, 2008.
Manuscript revised July 5, 2008.

[†]The authors are with the Department of Social Informatics, Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The author is with Language Grid Project, National Institute of Information and Communications Technology (NICT), Kyoto-fu, 619–0289 Japan.
[†††]The author is with Kyoto R&D Center, NPO Pangaea, Kyoto-shi, 600–8411 Japan.
a) E-mail: cho@ai.soc.i.kyoto-u.ac.jp

DOI: 10.1093/ietisy/e91–d.11.2638

*Examples of web-based content management applications include Pixiv (http://www.pixiv.net/), Flickr (http://flickr.com/), YouTube (http://youtube.com/), etc.

URLs to browse similar documents or browse documents in a top-down manner: in both cases, the aim is to expand users' browsing experience through tags. By contrast, our approach uses only tags (or interpretation word) and the frequency (or ratio) of the tags to assist users with the selection of more relevant pictograms: our goal is to provide users with an output that will aid the user to make more informed selection of the pictograms.

Our approach performs a priori categorization of pictogram interpretations and then calculates the *semantic relevance* to rank relevant pictograms for a given interpretation. We appropriate first level classifications in the Concept Dictionary of the EDR Electronic Dictionary [8] to define *pictogram categories* used for categorizing pictogram interpretations. We will show that categorized semantic relevance pictogram retrieval approach returns more stable result than not-categorized approach.

In the following section, a pictogram web survey for collecting interpretation words is briefly described, pictogram categories are explained, and characteristics in pictogram interpretation are clarified. Section 3 proposes semantic relevance measure and a priori categorization of interpretation words. Section 4 evaluates four different pictogram retrieval approaches using F_1 measure. Section 5 presents a pictogram sentence generator which leverages the proposed method. Finally, Sect. 6 concludes this paper.

2. Ambiguity in Pictogram Interpretation

2.1 Pictogram Web Survey

A pictogram web survey asking the meaning of 120 pictograms was conducted from October 1st, 2005 to November 7th, 2007 to collect free-answer English pictogram interpretation words and phrases. A total of 1,602 respondents living in the United States participated in the survey based on unique username–IP address pairs. For each pictogram, English interpretation words or phrases were first tallied according to unique interpretation word strings, and misspellings and single-occurrence interpretations were discarded. Single-occurrence interpretations indicate unique interpretations occurring only once across all 120 pictograms. Example of tallied pictogram interpretations is shown in Table 1. As shown, a pictogram can have various interpretations which include both similar and different-meaning words. For example, words like *strong*, *buff*, *healthy*, *muscular*, and *tough* all indicate a physical state of well-being whereas *hurt* refers to an injured state and *small* the size of an object. Action-related interpretations such as *flexing*, *workout*, and *exercise* are also given. When the focus shifts to the people depicted in the pictogram, the pictogram is interpreted as *strong man* or *body builder*. Or it can be interpreted as a kind of place such as *gym* or a specific object such as *muscle* or *muscles*.

Table 1 Example of tallied pictogram interpretations.

PICTOGRAM	WORD	FREQ.	RATIO
	strong	176	0.696
	muscle	33	0.130
	muscles	16	0.063
	strong man	7	0.028
	buff	3	0.012
	gym	3	0.012
	body builder	2	0.008
	flexing	2	0.008
	healty	2	0.008
	muscular	2	0.008
	tough	2	0.008
	workout	2	0.008
	exercise	1	0.004
	hurt	1	0.004
	small	1	0.004
	TOTAL	253	1.001

2.2 Five Pictogram Categories

One way to organize mixed interpretations containing both similar and different-meaning words is to group them into related perspectives. We use the Headconcept Dictionary and Concept Classification Dictionary of the EDR Electronic Dictionary [8] to categorize pictogram interpretation words[†].

The EDR Electronic Dictionary was developed for advanced processing of natural language by computers, and is composed of five types of dictionaries (Word, Bilingual, Concept, Co-occurrence, and Technical Terminology), as well as the EDR Corpus. The Concept Dictionary contains information on the approximately 410,000 concepts listed in the Word Dictionary and is divided according to information type into the Headconcept Dictionary, the Concept Classification Dictionary, and the Concept Description Dictionary. The Headconcept Dictionary describes information on the concepts themselves. The Concept Classification Dictionary describes the super-sub relations, i.e. inclusion relation, among the approximately 410,000 concepts [8]. We define *five pictogram categories* by appropriating the following five first level classifications defined in the Concept Dictionary:

(a) human or subject whose behavior (actions) resembles that of a human
(b) {matter} an affair
(c) event/occurrence

[†]SUMO ontology [9] was another candidate for categorizing pictogram interpretations, but we chose the EDR for three reasons: (1) we needed to handle both Japanese and English pictogram interpretations, and the EDR provides both English and Japanese headconcepts; (2) the first level classes located directly below SUMO ontology's Entity Class are Abstract Class and Physical Class, and these classes, we thought, were more abstract concepts than the first level classifications defined in the EDR; (3) EDR was specifically developed for natural language processing, therefore was more suitable to our research purpose which involves (pictogram) communication.

(d) location/locale/place
(e) time

For brevity, we abbreviate the pictogram category headings as (a) AGENT, (b) MATTER, (c) EVENT, (d) LOCATION, and (e) TIME; each maps to the aforementioned first level classifications respectively. Using the five pictogram categories, each pictogram interpretation word is categorized into appropriate pictogram categories through the following steps: first, concept identifier(s) of the interpretation word is obtained by matching the interpretation word string to the English headconcept string in the Headconcept Dictionary; then, the first level classification(s) of the concept identifier is obtained by climbing up the super-sub relations defined in the Concept Classification Dictionary. Note that since (i) more than one concept identifier may link to the same headconcept, and (ii) the Concept Classification Dictionary allows multiple inheritances, one interpretation word may be categorized into more than one pictogram category. More details on pictogram category acquisition can be found in [10].

2.3 Polysemous Interpretation

Table 2 shows the interpretation words of four pictograms obtained from the web survey. The first row shows pictogram numbers; the second row shows pictogram images (PICT.); the third row lists interpretation words according to the descending order of frequency (LIST OF INTERPRETATION WORDS); and the bottom two rows show underlined interpretation words categorized into two pictogram categories, AGENT and EVENT. Note that the list of interpretation words in pictograms (2), (3), and (4) are partial lists.

Categorizing the interpretation words into pictogram categories elucidates two key aspects of polysemy in pictogram interpretation. Firstly, interpretations that spread across different categories lead to different perspectives in interpretations. For example, Table 2 AGENT row includes words like *singer(s)*, *choir* and *chorus* which describe certain kind of people while EVENT row includes words like *singing*, *happy*, *loud*, and *reading* which describe ongoing actions or states; the two word groups contain different perspectives.

Secondly, while interpretation words placed within the same pictogram category may contain related words such as *singer(s)*, *choir* and *chorus* in the AGENT row, or *talking*, *speaking*, and *talk* in column (3) (they can be categorized into the EVENT category), different-meaning words sometimes coexist within the same category. For example, column (3) contains *talking*, *praying*, *thinking*, *reading*, and *singing* which are different EVENT category words, and column (4) contains *workers*, *teachers*, and *singers* which are different AGENT category words. To summarize the first and second findings, it can be said that polysemy in pictogram interpretation is generally observed 'across' different categories, but varied interpretations may also be found 'within' the same category.

Table 2 Polysemous interpretations within each pictogram (each column), shared interpretations across pictograms (underlined words), and shared interpretations categorized into two pictogram categories (bottom two rows).

	(1)	(2)	(3)	(4)
PICT.				
LIST OF INTERPRETATION WORDS	singing sing singer music happy song choir chorus musical loud	excited happy surprised yelling screaming yell shout angry surprise loud singing	talking praying thinking speaking lonely speech pray talk reading singing speaker	jobs family reading careers workers teachers choir singing singers chorus school
AGENT	singer choir chorus	-	-	choir singers chorus
EVENT	singing happy loud	happy loud singing	reading singing	reading singing

When a pictogram having polysemous interpretations is used in communication, there exists a possibility that the sender and receiver might interpret the same pictogram differently. For instance, in the case of pictogram (4), it can be interpreted differently as *teachers* and *singers* by the sender and receiver respectively. One way to assist the sender to select a pictogram that has a higher chance of conveying the intended meaning is to display possible interpretations of the pictogram. If various possible interpretations are presented, the sender can speculate on the receiver's interpretation before selecting a pictogram. For example, if the sender knows a priori that pictogram (4) can be interpreted as both *teachers* and *singers*, he or she can guess ahead that it might be interpreted differently by the receiver, and avoid choosing the pictogram. The displaying of possible pictogram interpretations is the first pictogram selection assistance we propose in tackling the issue of polysemy or one-to-many correspondence in pictogram-to-pictogram interpretations.

2.4 Shared Interpretation

A single pictogram may contain various interpretations, but these interpretations are not necessarily exclusive to one pictogram; sometimes two or more pictograms may share the same interpretation(s). Underlined words in Table 2 are such interpretations shared by more than one pictogram: for example, all four pictograms share the word *singing*; pictograms (1) and (4) share *singer(s)*, *choir*, and *chorus*; pictograms (1) and (2) share *happy* and *loud*; and pictograms (3) and (4) share the word *reading*.

The fact that multiple pictograms can share common interpretation word implies that each one of those pictograms can be interpreted as such. The degree to which each is interpreted as the shared interpretation, however, may vary according to the pictogram. For instance, all four pictograms in Table 2 can be interpreted as *singing*, but pictogram (4) can be also interpreted as *reading*. Furthermore, if we look at the words in the AGENT category, we see that pictogram (1) mainly contains description of people engaged in the act of singing (*singer*, *choir*, *chorus*) whereas pictogram (4) contains more varied people descriptions such as *family*, *workers*, and *teachers*.

Suppose two people A and B each select pictogram (1) and (4) respectively to send a message about a singing activity to person C. Upon receiving the pictogram message however, C may interpret A's message as "singing activity" while interpret B's message as "reading activity." Even though A and B both intend to convey "singing activity" to C, it may not be the case that C will interpret both pictograms likewise; this is because the degree of the shared interpretation, which in this case is *singing*, may vary across different pictograms. Such degree difference, we think, is affected by both the probability of the shared interpretation in each pictogram and the remaining interpretations within each pictogram. For example, the probability of the word *singing* in Table 2 pictograms (1) and (4), and the remaining interpretations such as *musical* and *singer* in pictogram (1) versus *reading* and *teacher* in pictogram (4), both affect how strong pictograms (1) and (4) can be interpreted as *singing*.

When two or more pictograms share the same interpretation I, the degree to which each pictogram may be interpreted as I may vary. If the degree of interpretation I for each pictogram is known, selecting the pictogram with the greatest degree of I will increase the chance of conveying I compared to other candidate pictograms. Hence, one way to assist pictogram selection among multiple pictograms sharing the same interpretation is to rank those pictograms according to the degree of relevancy to a given interpretation. In order to rank pictograms according to the relevancy of certain interpretation, some kind metric which measures the relevancy of a pictogram to a given interpretation is needed; to this end, we propose a *semantic relevance measure*. The calculation of semantic relevance and the ranking of pictograms according to the interpretation relevancy is the second pictogram selection assistance we propose in tackling the issue of many-to-one correspondence in pictograms-to-pictogram interpretation. We describe the semantic relevance measure next.

3. Categorical Semantic Relevance

3.1 Semantic Relevance Measure

We assume that pictograms each have a list of interpretation words and ratios as the one given in Table 1. Each unique interpretation word has a frequency, and each word frequency indicates the number of people who answered the pictogram to have that interpretation. The ratio or the probability of an interpretation word, which can be calculated by dividing the word frequency by the total word frequency of that pictogram, indicates how much support people give to that interpretation. For example, in the case of the pictogram in Table 1, it can be said that more people support *strong* (176 out of 253) as the interpretation for that pictogram than *hurt* (1 out of 253). The higher the ratio or the probability of a specific interpretation word in the pictogram, the more that pictogram is accepted by people for that interpretation. We define *semantic relevance measure* of a pictogram to be the measure of relevancy between a word query and interpretation words of a pictogram.

Let $w_1, w_2, \ldots, w_n$ be interpretation words of pictogram e. Let the probability of each interpretation word in a pictogram to be $P(w_1|e), \ldots, P(w_n|e)$. For example, the probability of the interpretation word *strong* for Table 1 pictogram can be calculated as $P(strong|Pictogram_{Table\ 1}) = 176/253$. Then the simplest expression that assesses the relevancy of a pictogram e in relation to a word query w_i can be defined as follows:

$$P(w_i|e) \tag{1}$$

This probability, however, does not take into account the similarity of interpretation words. For instance, when "strong" is given as query, pictograms having similar interpretation words like *brawny* or *stout*, but not *strong*, fail to be measured as relevant when only the probability is considered. To solve this, we need to define some kind of similarity, or $similarity(w_i, w_j)$, between interpretation words. Using the similarity, we can define the *semantic relevance measure* or $SR(w_i, e)$ as follows:

$$SR(w_i, e) = \sum_j P(w_j|e) similarity(w_i, w_j) \tag{2}$$

There are several similarity measures. We draw upon the definition of similarity given by Lin [11] which states that similarity between A and B is measured by the ratio between the information needed to state the commonality of A and B and the information needed to fully describe what A and B are. Here, we calculate the similarity of w_i and w_j by counting how many pictograms contain certain interpretation words. When there is a pictogram set E_i having an interpretation word w_i, the similarity between interpretation words w_i and w_j can be defined as follows:

$$similarity(w_i, w_j) = |E_i \cap E_j| / |E_i \cup E_j| \tag{3}$$

$|E_i \cap E_j|$ is the number of pictograms having both w_i and w_j as interpretation words. $|E_i \cup E_j|$ is the number of pictograms having either w_i or w_j as interpretation word. Based on (2) and (3), the *semantic relevance* or the measure of relevancy to return pictogram e when a word w_i is input as query can be calculated as follows:

$$SR(w_i, e) = \sum_j P(w_j|e) |E_i \cap E_j| / |E_i \cup E_j| \tag{4}$$

The calculated semantic relevance values fall between one and zero, which denotes that either a pictogram is completely relevant to the interpretation (query) or completely irrelevant. Using the semantic relevance values, pictograms can be ranked from very relevant (value close to 1) to not so relevant (value close to 0). As the value nears zero, pictograms become less relevant; hence, a cutoff point is needed to discard the less relevant pictograms. Setting an ideal cutoff point that satisfies all word query and pictogram interpretations is difficult, since all words contained in a pictogram, regardless of how great or small each interpretation word is related to the query, influence the semantic relevance calculation. For example, let's say that we want to find a pictogram which can convey the meaning "workout." Pictogram in Table 1 could be a candidate since it contains *workout* with a ratio of 0.008. When the semantic relevance value is calculated, however, the equation takes into account not only the interpretation word matching the query, but all the remaining interpretation words including *strong*, *muscle(s)*, *gym*, *body builder* and so forth. So, one way to remedy the dispersion of interpretation is to select a set of interpretation words more related to the query, and use those selected words in the semantic relevance calculation to reduce the effect of less-related interpretation words affecting the calculation. With this prediction, we propose a semantic relevance calculation on categorized interpretations.

3.2 Categorizing the Pictogram Interpretations

Pictogram interpretation words are categorized into the five pictogram categories described in Sect. 2.2. Note that one interpretation word may be categorized into multiple pictogram categories since a word may link to multiple concept identifiers via the same headconcept or via multiple inheritances. For example, in the case of the word (headconcept) *park*, three kinds of pictogram categories are obtained repeatedly: LOCATION category six times, MATTER category five times, and EVENT category four times. In such case of multiple category acquisition, we use all categories since we cannot accurately guess on the single correct category intended by each respondent who participated in the web survey.

3.3 Weighting the Pictogram Interpretations

Although we cannot correctly decide on the single, intended category of a word, we can calculate the ratio of the pictogram category of each word. For example, in the case of the word *park*, the LOCATION category has the most number of repeated categories (six). Next is the MATTER category (five) followed by the EVENT category (four). We can utilize such category constitution by calculating the ratio of the repeated categories and assigning the ratio as weights to the word in a given category. For example, the word *park* can be assigned to LOCATION, MATTER and EVENT category, and for each category, weights of 6/15, 5/15 and 4/15 can be assigned to the word. Consequently, the *major category* of the interpretation word *park* will be LOCATION.

3.4 Ranking the Result

Applying the semantic relevance calculation to categorized interpretations will return five *categorical semantic relevance values* for each pictogram. We take the highest categorical semantic relevance value and compare it with the cutoff point to determine whether the pictogram is relevant or not. Once the relevant pictograms are selected, the selected pictograms are then sorted according to the semantic relevance value of the query's major category. For example, if the query is "park", then the relevant pictograms are first selected using the highest categorical semantic relevance value of each pictogram, and once the relevant pictograms are selected, the pictograms are ranked according to the categorical semantic relevance value of the query's major category, which in this case is the LOCATION category. The resulting list of pictograms is a ranked list of pictograms starting with the most relevant pictogram on top.

3.5 Prototype Implementation

We implemented a prototype web-based pictogram retrieval system which returns a list of relevant pictograms in the descending order of the query's major category's semantic relevance values when a word query is given as input. The categorized and weighted pictogram interpretation words for 120 pictograms were given to the system as data to calculate the categorical semantic relevance values. Figure 1 shows a list of retrieved pictograms for the query "slide." Note that the retrieved pictograms are sorted according to the EVENT category's semantic relevance values since the major category of the query "slide" is EVENT. When the pictograms are ranked according to the LOCATION category's semantic relevance values, however, the ranking changes with the fourth pictogram with the highest LOCATION value (0.29112) jumping to the top. This difference in the categorical semantic relevance value will be utilized in the pictogram sentence generator described in Sect. 5.

The prototype system implements the two design principles discussed in Sects. 2.3 and 2.4 which deals with one-to-many (i.e. polysemous) and many-to-one (i.e. shared) relationship between pictogram and pictogram interpretations. Evaluation of the proposed method is described next.

4. Evaluation

4.1 Comparison of the Four Approaches

Three pictogram retrieval approaches that singly uses or combines the semantic relevance measure, word categorization, and word weighting were evaluated. The baseline for comparison was a simple string match of the word query to the pictogram interpretation words with probabilities greater than the cutoff point. This is the same as selecting pictograms with $P(w_j|e) > \textit{cutoff point}$ where w_j equals the

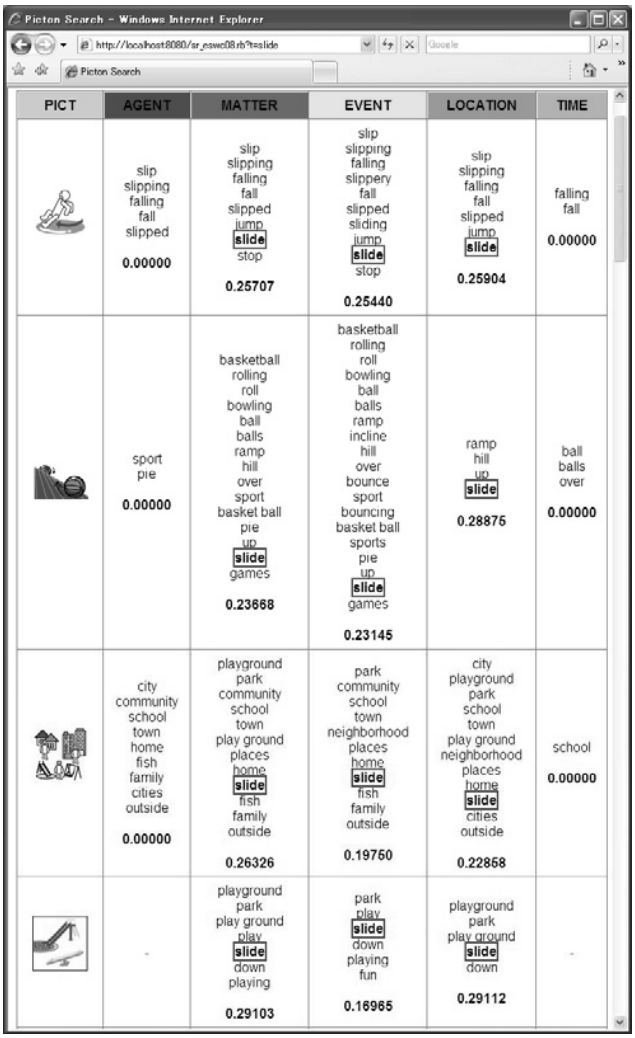

Fig. 1 A screenshot of a prototype web-based pictogram retrieval system which uses the categorized and weighted semantic relevance approach. Results for the query "slide" is displayed.

query. A relevant pictogram set was constructed by five human judges, and retrieval tasks were performed using the four approaches: (1) baseline string match approach, (2) not-categorized semantic relevance approach, (3) categorized semantic relevance approach, and (4) categorized and weighted semantic relevance approach.

4.2 Relevant Pictogram Set Construction

Five human judges were employed in the construction of a relevant pictogram set which consists of 188 pictogram interpretation words and a ranked list of relevant pictograms for each word. The judges were all undergraduate students and they were paid for their tasks. The relevant pictogram set was constructed through the following steps:

[STEP 1] Collecting Human Assessment Data: A questionnaire containing 188 pictogram interpretation words[†] with candidate pictograms[††], each listing all interpretation words (similar to the second column in Table 1)[†††], was given to the five human judges, and for each interpretation word, the human judges were asked to (i) judge whether each candidate pictogram could be interpreted as the given word (i.e. judged either as relevant or not relevant), and (ii) if judged as relevant, write down the ranking among the relevant pictograms.

[STEP 2] Judging and Ranking Relevant Pictograms: The five judges' assessment data were averaged and variances were calculated to select and rank relevant pictograms for each interpretation word. If three or more people judged the pictogram to be relevant, the pictogram was selected as relevant. Otherwise, the pictogram was discarded. Average rankings among the selected pictograms were calculated based on the rankings given by the human judges; if average rankings were the same among two or more pictograms, variances were calculated to give higher ranking to the pictogram with lower variance. As a result, a ranked relevant pictogram set for 188 words were created and used in the evaluation.

4.3 Precision, Recall, and F_1 Measure

The mean precision, mean recall, and F_1 measure [12] of 188 retrieval tasks on the four pictogram retrieval approaches were calculated using nine different cutoff points from 0.1 to 0.5 with 0.05 intervals. Figure 2 shows F_1 measure, and Tables 3 and 4 respectively show the mean precision and mean recall: SR-WCAT indicates categorized and weighted semantic relevance approach; SR-CAT indicates categorized semantic relevance approach; SR-NOCAT indicates not-categorized semantic relevance approach; and STR-MATCH indicates the baseline string match approach. Note that the mean precision values were calculated using the valid tasks that returned at least one result. For example, in the case of cutoff value 0.5, only 9 retrieval tasks returned at least one pictogram for the STR-MATCH approach; hence, the mean precision of the STR-MATCH ap-

[†]There were initially a total of 903 unique pictogram interpretation words for 120 surveyed pictograms which could be used as word queries for the retrieval task. We first performed retrieval tasks with these 903 words using the four approaches to eliminate 399 words that returned the same result for all four approaches, since these words would be ineffective in discerning the four approaches' retrieval performance. Another 216 words which returned the same results for the three semantic relevance approaches were eliminated. 288 words remained as a result. Among the 288 words, words having more than nine candidate pictograms, similar words (e.g. *hen*, *rooster*), singular/plural words (e.g. *girl*, *girls*), and varied tenses (e.g. *win*, *winning*) were eliminated leaving 188 words to be judged for relevancy. The constitution of major pictogram categories in the 903 words and 188 words were:
-903:[Agent,10%],[Matter,24%],[Event,61%],[Location,2%],[Time,3%]
-188:[Agent, 9%],[Matter,28%],[Event,50%],[Location,9%],[Time,5%]

[††]Candidate pictograms contain given interpretation.

[†††]The probability of each interpretation word was not displayed in the questionnaire, but was used to rank the words in the descending order.

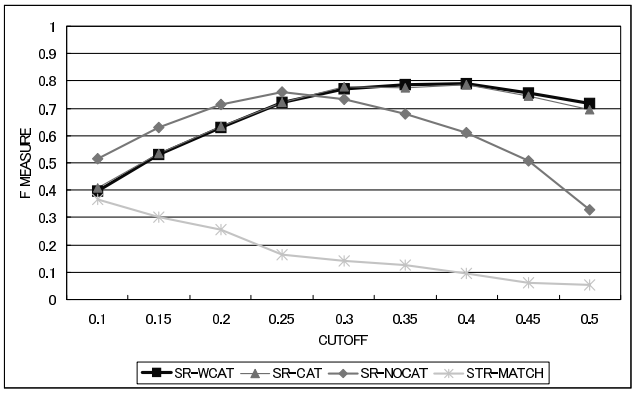

Fig. 2 Comparison of F_1 measure graph of four approaches.

Table 3 Mean precision of four approaches at different cutoffs.

Cutoff	SR-WCAT	SR-CAT	SR-NOCAT	STR-MATCH
0.10	0.24850	0.25810	0.34883	0.98056
0.15	0.36089	0.36467	0.46397	0.99275
0.20	0.46108	0.46512	0.57565	1.00000
0.25	0.57463	0.57928	0.67917	1.00000
0.30	0.65529	0.66786	0.73870	1.00000
0.35	0.70685	0.70442	0.79100	1.00000
0.40	0.73910	0.74880	0.84497	1.00000
0.45	0.76704	0.76979	0.87760	1.00000
0.50	0.78753	0.81036	0.89655	1.00000

Table 4 Mean recall of four approaches at different cutoffs.

Cutoff	SR-WCAT	SR-CAT	SR-NOCAT	STR-MATCH
0.10	1.00000	1.00000	0.99867	0.22615
0.15	1.00000	0.99823	0.97442	0.17766
0.20	0.99493	0.98980	0.94174	0.14752
0.25	0.96226	0.95713	0.86184	0.08901
0.30	0.94125	0.93784	0.72376	0.07704
0.35	0.88712	0.86527	0.59734	0.06640
0.40	0.84705	0.82768	0.47810	0.05044
0.45	0.74785	0.72214	0.35887	0.03183
0.50	0.65888	0.60657	0.20222	0.02739

proach was calculated using only those 9 tasks. Note that gain in retrieval performance is achieved through semantic relevance and word categorization, but minimal gain is obtained through word weighting.

5. Discussions

We see in Fig. 2 that a broader cutoff range between 0.24 and 0.5 is obtained by the categorized approach for F_1 measure greater than 0.7 (SR-WCAT & SR-CAT) whereas the not-categorized approach has a more steeper curve with narrower cutoff range between 0.19 and 0.33 (SR-NOCAT). The wider range of stable F_1 measures given by the categorized approach owes to a priori grouping of the interpretation words into related perspectives; this enables targeted semantic relevance calculation on words more related to the query and related to each other leading to the improvement in recall without damaging precision. This is confirmed in Tables 3 and 4: in Table 4, the recall range of SR-WCAT and SR-CAT is tighter with the range approximately between 0.6 and 1.0 whereas SR-NOCAT is broader with recall range approximately between 0.2 and 1.0; meanwhile in Table 3, the precision range of all three approaches, SR-WCAT, SR-CAT, and SR-NOCAT, are similar with SR-WCAT and SR-CAT approximately in between 0.25 and 0.8, and SR-NOCAT in between 0.35 and 0.9.

The fact that no significant performance gain was obtained through category weighting of the words should be discussed. The categorical semantic relevance values of SR-WCAT and SR-CAT did not differ greatly in most cases of the retrieved results (in general, SR-WCAT had slightly higher values than SR-CAT). As a result, the same set of pictograms was retrieved for both approaches except in those cases where the two categorical semantic relevance values branched at the cutoff point. Analyzing the exceptional cases where large value differences between the two approaches were observed revealed that the combination of large category weight on the query word together with small category weight on the surrounding interpretation words triggered a drastic change in the interpretation word ratio, causing the categorical semantic relevance value of SR-WCAT to increase drastically. Such an exceptional case was rarely observed, however, and in most cases the constitution of the interpretation word ratio in SR-WCAT and SR-CAT were very similar. This is why the two approaches exhibited very similar retrieval performances.

Our method can be applied to various image management applications such as clipart search systems or online photo-sharing systems as long as the images are labeled with descriptive tags, and that those tags have frequencies; but the benefits of categorization can be more fully enjoyed through a novel application which generates a pictogram-mixed sentence. We will call it a *pictogram sentence generator*. A pictogram sentence generator is a parser-like application which takes a text sentence as input and outputs a pictogram-mixed sentence. The generator first parses the text sentence to generate a parse tree, and then takes the lemma of the word in the tree to use it as a word query to search for the most relevant pictogram to replace the word. The pictogram retrieval system introduced in Sect. 3.5 is utilized in the search process, but instead of ranking the retrieved result using the major pictogram category of the query, the generator specifies which pictogram category to emphasize, i.e. which categorical semantic relevance value should be selected and ranked. We explain this using two examples. Suppose we want to convert the following two sentences, both containing the word "slide", into a pictogram-mixed sentence:

(1) John likes to slide down the hill.

(2) John played on the slide.

It is obvious to humans that the word "slide" is used differently in the two sentences, but a simple string match by a machine will find no difference. One way to allow a machine to discern this kind of usage difference is to provide the machine with semantic role information. Recent advances in semantic role labeling technology have realized

fairly accurate automatic labeling of semantic roles[†], and currently several semantic role labelers[††] have been implemented. Some labelers uses semantic roles defined in the Proposition Bank [13], and the numbered arguments in the frameset are aligned to VerbNet [14] thematic roles. If we can map these thematic roles, adjuncts labels, verb information to the first, second, and third level classifications in the Concept Dictionary of the EDR Electronic Dictionary, then once the semantic roles are identified, we can obtain the first level classification in the EDR (i.e. the most appropriate pictogram category) that can be used for ranking the retrieved pictograms.

Going back to the two example sentences, the semantic role labeler will output "verb" as label for the word "slide" in sentence (1) and "AM-LOC location" for "slide" in (2); then, we can map "slide: verb" to `30f83e action/act` in the Concept Dictionary to obtain the EVENT category whereas map "slide: location" to `30f751 location/locale/place` to obtain the LOCATION category. When the retrieved pictograms are ranked using the acquired pictogram categories, the generator will output Fig. 1's top-most pictogram showing a person sliding for the word "slide" in sentence (1) and the bottom-most pictogram showing a playground slide for the word "slide" in (2).

The number of the human judges participated in the evaluation experiment given in this paper were few and the age group was limited, and so a greater number of human judges encompassing a wide range of age groups should be incorporated, and more diverse queries should be used for performance evaluation to accommodate real-world usage. Moreover, a reliability metric to mitigate the overrated calculation of single word categorical semantic relevance value should be defined to improve precision in the future.

6. Conclusion

Polysemous and shared pictogram interpretation can lead to ambiguity in pictogram interpretation, which can cause misunderstanding in communication using pictograms. To retrieve pictograms that can better convey the intended meaning, we proposed a method of selecting and ranking relevant pictograms which are more likely to be interpreted as intended. We proposed a categorical semantic relevance measure, which calculates how relevant a pictogram is to a given interpretation in terms of a pictogram category. The measure defines the probability and similarity measurement of categorized pictogram interpretations. Five pictogram categories used for categorizing pictogram interpretation words were defined using the Concept Dictionary of the EDR Electronic Dictionary. Three semantic relevance approaches, (i) not-categorized semantic relevance approach, (ii) categorized approach, and (iii) categorized and weighted approach, were evaluated using five human judges and 188 queries, and the categorized approaches showed more stable performance than the not-categorized approach.

[†]The labeling performances (F_1 measure) of the CoNLL-2005 Shared Task's top-ranking participants were in the high seventies.

[††]An online semantic role labeler by UIUC can be found at http://l2r.cs.uiuc.edu/~cogcomp/srl-demo.php

Acknowledgements

This research was partially supported by the International Communications Foundation and the Global COE Program "Informatics Education and Research Center for Knowledge-Circulating Society." We thank Dr. Daisuke Kawahara at NICT for the valuable discussions on semantic role labeling. *All pictograms presented in this paper are copyrighted material, and their rights are reserved to NPO Pangaea.*

References

[1] T. Takasaki, "Design and development of a pictogram communication system for children around the world," in Intercultural Collaboration, ed. T. Ishida, S.R. Fussell, and P. Vossen, LNCS, vol.4568, pp.193–206, Springer, 2007.
[2] B.R. Baker, "Minspeak, a semantic compaction system that makes self-expression easier for communicatively disabled individuals," Byte, vol.7, no.9, pp.186–202, 1982.
[3] C. Beardon, "Discourse structures in iconic communication," Artif. Intell. Rev., vol.9, pp.189–203, 1995.
[4] A. Marcus, "Icons, symbols, and signs: Visible languages to facilitate communication," Interactions, vol.10, no.3, pp.37–43, 2003.
[5] P.A. Kolers, "Some formal characteristics of pictograms," American Scientist, vol.57, pp.348–363, 1969.
[6] M. Aurnhammer, P. Hanappe, and L. Steels, "Augmenting navigation for collaborative tagging with emergent semantics," ISWC '06 Proc. 5th Int'l Semantic Web Conf., LNCS, vol.4273, pp.58–71, Springer, 2006.
[7] R. Li, S. Bao, B. Fei, Z. Su, and Y. Yu, "Towards effective browsing of large scale social annotations," WWW '07 Proc. 16th Int'l World Wide Web Conf., pp.943–952, ACM Press, 2007.
[8] National Institute of Information and Communications Technology (NICT), The EDR Electronic Dictionary. Available at: http://www2.nict.go.jp/r/r312/EDR/
[9] I. Niles and A. Pease, "Towards a standard upper ontology," FOIS '01 Proc. 2nd Int'l Conf. on Formal Ontology in Information Systems, 2001.
[10] H. Cho, T. Ishida, T. Takasaki, and S. Oyama, "Assisting pictogram selection with semantic interpretation," ESWC '08 Proc. 5th European Semantic Web Conf., LNCS, vol.5021, pp.65–79, 2008.
[11] D. Lin, "An information-theoretic definition of similarity," ICML '98 Proc. 15th Int'l Conf. on Machine Learning, pp.296–304, 1998.
[12] C. van Rijsbergen, Information Retrieval, Butterworths, 1979.
[13] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," Computational Linguistics, vol.31, no.1, pp.71–105, 2005.
[14] K.K. Schuler, "VerbNet: A broad-coverage, comprehensive verb lexicon," Ph.D. Thesis, University of Pennsylvania, 2005.

Heeryon Cho is a PhD candidate at the Department of Social Informatics, Kyoto University. She received B.A. in Mass Communication from Yonsei University, Seoul, Korea in 1995, and Master of Informatics from Kyoto University, Kyoto, Japan in 2005.

Toru Ishida is a professor in the Department of Social Informatics, Graduate School of Informatics, Kyoto University from 1993, a visiting professor of Shanghai Jiao Tong University from 2002. He has been working on Digital City and Language Grid projects. He is a fellow of IEICE, IPSJ, and IEEE.

Satoshi Oyama is an assistant professor in the Department of Social Informatics, Graduate School of Informatics, Kyoto University. He received the B.Eng., M.Eng., and Ph.D. degrees from Kyoto University in 1994, 1996, and 2002, respectively. He was a research fellow of the Japan Society for the Promotion of Science from 2001 to 2002. He was a visiting assistant professor in the Department of Computer Science at Stanford University from 2003 to 2004.

Rieko Inaba is a research fellow at the Language Grid Project, Knowledge Creating Communication Research Center in National Institute of Information and Communication Technology since 2006. She received B.A., M.S., and PhD in Science from Department of Mathematical and Physical Science, Japan Women's University in 1998, 2000, and 2003 respectively.

Toshiyuki Takasaki Toshiyuki Takasaki is a vice president of NPO Pangaea and a GCOE researcher in the Department of Social Informatics, Kyoto University. He received B.A. at Precision Engineering Department, University of Tokyo, and M.A. at Graduate School of Frontier Science, University of Tokyo.