Toward Communicating Simple Sentences Using Pictorial Representations

Rada Mihalcea
Computer Science Department, University of North Texas
rada@cs.unt.edu

Chee Wee Leong
Computer Science Department, University of North Texas
cl0198@cs.unt.edu

August 26, 2008

Abstract. This paper addresses and evaluates the hypothesis that pictorial representations can be used to effectively convey simple sentences across language barriers. The paper makes two main contributions. First, it proposes an approach to augmenting dictionaries with illustrative images using volunteer contributions over the Web. The paper describes the PicNet illustrated dictionary, and evaluates the quality and quantity of the contributions collected through several online activities. Second, starting with this illustrated dictionary, the paper describes a system for the automatic construction of pictorial representations for simple sentences. Comparative evaluations show that a considerable amount of understanding can be achieved using visual descriptions of information, with evaluation figures within a comparable range of those obtained with linguistic representations produced by an automatic machine translation system.

Keywords: text-to-picture synthesis, illustrated dictionaries, augmentative and alternative communication

1. Introduction

According to recent studies (Ethnologue, 2005; Gibbs, 2002) there are about 7,000 languages spoken worldwide. Currently, only about 15–20 languages can take advantage of the benefits provided by machine translation, and even for this subset of languages, the automatically produced translations are not error free and their quality lags behind the human expectations.

In this paper, we explore the use of **pictorial representations** as a means for conveying information across language barriers. Regardless of the language they speak, people share almost the same ability to understand the content of pictures. For instance, speakers of different languages have a different way of referring to the concept of *apple*. Instead, a picture can be understood by all people in the same way, replacing the multitude of linguistic descriptions with one, virtually universal representation.

In addition to enabling communication across languages, the ability to encode information using pictorial representations has other benefits, such as language learning for children or for those who study a foreign language (Carney and Levin, 2002), communication with preliterate or non-literate people (Medhi *et al.*, 2006), or language understanding for people with language disorders (Behrmann and Byng, 1992; Alm *et al.*, 2002).

This paper describes a system for the automatic construction of pictorial representations for simple sentences, and performs an initial assessment of the amount of understanding that can be achieved using visual descriptions of information. An example of the pictorial representations that we target is shown in Figure 3(a).

The main idea behind the experiments reported in this paper is to use sequences of pictures to form the gist of a sentence. We do so by replacing the concepts in the sentence with their pictorial representations. In the terminology used in this paper, a concept is defined as “an abstract idea or mental symbol, inferred or derived from specific instances.” While concepts are typically associated with a corresponding representation in a language (words), in our experiments we try to associate concepts with a pictorial representation. As with words, this association is only an approximation of the meaning of the concept.

Note that an image can and does bring to mind several concepts. For instance, a picture of a rabbit contains the concept of “rabbit,” but also those of “fur,” “ears,” etc. However, when one sees the picture of a rabbit, the first concept that comes to mind is that of “rabbit,” and thus this is a more salient concept among the alternatives. This is similar to the scenario when one sees the word “rabbit,” which will first bring to mind the concept of “rabbit,” which is more salient than other concepts that can be inferred from the same lexical representation, such as “animal” or “words starting with r”.

Through our experiments, we evaluate if the interpretation of a sentence using visual representations (pictures) remains consistent with the linguistic representations (words) in the original sentence.

There are of course limitations inherent to the use of such visual representations of information. First, there are complex meanings that cannot be conveyed through pictures in a straight-forward manner, as in e.g. “*An inhaled form of insulin won federal approval yesterday,*” which requires the more advanced representations that are more easily encoded in a language. In particular, linguistic representations have hierarchical and relational properties, which render them more powerful in terms of expressiveness. On the other hand, pictorial representations may introduce “noise,” as the peculiarities of objects in a picture tend to cause difficulty in the standardization of what we mean with that particular picture.

Second, there is a large number of concepts that have a level of abstraction that prohibits a visual representation, e.g., *politics* or *regenerate*. While symbols can also be used to encode abstract concepts, the representation of such concepts needs to be learned by the user, a process that most people have already accomplished in relation to language, but requires additional effort when it comes to pictorial representations.

Finally, it has been shown that cultural differences may result in varying levels of understanding for certain concepts (Nakamura *et al.*, 1998; Huer, 2000; Nigam, 2003). For instance, the prototypical image for *house* may be different in Asian countries as compared to countries in Europe. Therefore, the pictorial representations of concepts may not be understood in the same way by speakers of different languages.

While we acknowledge all these limitations and difficulties, we attempt to take a first cut at the problem, and evaluate the amount of understanding for simple sentences when "represented through pictures," as compared to the typical linguistic representations. Note that we do not attempt to represent complex states or events (e.g., temporal markers, idioms with metaphorical meaning), nor do we attempt to communicate their attributes (adjectives, adverbs). Instead, we focus on generating pictorial representations for simple sentences, using visual descriptions for basic concrete nouns and verbs (as defined in (Coltheart, 1981)), and we evaluate the amount of understanding that can be achieved with these simple visual descriptions as compared to their linguistic alternatives.

The paper is organized as follows. We start by overviewing previous related work on pictorial and symbolic representations. We then describe the construction of the PICNET illustrated dictionary, and show how this knowledge-base can be use to build a system for generating pictorial representations for simple sentences. We evaluate the quality of the representations under three different scenarios, followed by a discussion of the results. We conclude the paper with suggestions for future work.

2. Related Work

Early research efforts in cognitive science and psychology (Potter *et al.*, 1986) have shown that a picture can successfully replace a noun in a rapidly presented sentence, without any impact on the interpretation of the sentence, nor on the speed of understanding, suggesting that the human representation of word meanings is based on a conceptual system which is not tied to a given language. These findings have recently found support in cross-cultural studies that showed that children from different countries, not speaking each other's language, were able to communicate about children stories just by using drawings and pictures (Alburo *et al.*, 2005).

The work most closely related to ours is the WordsEye project (Coyne and Sproat, 2001), which targets the generation of scenes starting with an input text. The system is gradually building a scene by adding objects identified in a text; it is meant as a support tool for graphic designers, and not necessarily as a communication system. In fact, although WordsEye's database consists of thousands of object models, the system works only for descriptive sentences of collated objects, and so it cannot generate scenes for sentences such as "The house has four bedrooms and one kitchen" (where it has only "house" as a picturable object).

Another closely related project is Symbolate (Symbolate, 2008), which is an application used primarily for educational purposes. Symbolate adds symbols to the words in a sentence, with the goal of helping users recognize words by using pictures as clues. The assumption is that both words and pictures are available, and thus the interpretation of sentences in Symbolate is not based exclusively on pictorial representations, as it is in our work, but rather on pictorial and linguistic representations together.

Other related projects along similar lines are SPRINT (Yamada *et al.*, 1992), where geometric models are created from natural language descriptions of a scene, using spatial constraints extracted from the text; Put (Clay and Wilhelms, 1996), which identifies the placement of objects in a scene using an interactive natural language interface; and CarSim (Johansson *et al.*, 2005), which converts narratives about car accidents into 3D scenes by using techniques for information extraction coupled with a planning and a visualization modules. More recently, the text-to-picture TTP system for augmentative communication (Zhu *et al.*, 2007) was used to synthesize a picture from natural language text by finding the important concepts in the text and merging the pictorial representations of these concepts.

Work has also been done on the design of iconic symbol sets for augmentative and alternative communication (AAC) for people with physical limitations or speech impediments, with iconic keyboards that can be touched to produce a voice output for communication augmentation (Chang *et al.*, 1992). In a comparative evaluation across five AAC symbol systems and sets (Bliss symbols, Picsyms, PIC, PCS, Rebus), it was found that the ability to rate the match between symbols and meanings (also referred to as translucency) can differ significantly from a symbol set to another (Bloomberg *et al.*, 1990). Other comparative studies, e.g., (Mizuko, 1987; Musselwhite and Ruscello, 1984), have also found significant differences between various symbol sets in terms of ability of users to guess the meaning of the symbols (guessability) and the ability to find the best match between a given meaning and several candidate symbols (transparency).

Work on AAC has also been concerned with multi-linguality and cross cultural differences. For instance, a multi-lingual augmentative communication system was found helpful in giving a multilingual capability to non-speaking people or to people who do not speak a foreign language (Alm *et al.*, 2002). In related work, cross cultural studies reported in (Haupt and Alant, 2003) have found significant differences in the interpretation of symbolic sets by people from different cultural backgrounds (e.g., American and Zulu participants). Moreover, studies also found that the level of education can impact the correctness of the interpretation, with older, higher educated participants having better performance (Evans *et al.*, 2006), and illiterate or low literate participants having low performance (Hanson and Hartzema, 1995).

Also related to some extent is the work done in visual programming languages (Boshernitsan and Downes, 1997), where visual representations such as graphics and icons are added to programming languages to support visual interactions and to allow for programming with visual expressions.

Finally, a significant amount of research work has been done in automatic image captioning, e.g., (Barnard and Forsyth, 2001; Pan *et al.*, 2004). In particular, PICNET relates to two other Web-based projects that try to bridge concepts and images through image annotations.

The first project, called the ESP game (von Ahn and Dabbish, 2004), is an online system that collects labels associated with images. The system is set up as a game, where the goal is to assign as many labels as possible to a given image. When two players concurrently assign a label, the label is considered correct and stored in the set of tags associated with the image. Unlike PICNET, which targets the assignment of pictures to words, the ESP game collects words (labels) for pictures. Most of the pictures labeled by the ESP game consist of entire scenes, which often refer to several concepts. For instance, an ESP-annotated image could have the following label assignments: *car*, *person*, *tree*, *house*, *road*. While it is possible that a multiple-object scene be used to describe a unique concept, the ESP game does not have any constraints concerning the number of concepts associated with an image, and there are often multiple salient concepts associated with an image in the ESP database.

Another related project is the Google Image Quiz.[1] Provided a set of images returned by a search performed against the Google Image search engine, the goal is to guess the keyword that was used in the search. Similar to the ESP game, the Google Image Quiz assigns labels (words) to images, and not images to words, and thus it can often be the case that an image will refer to several salient concepts that are associated with it.

Although related in their goal of connecting images and concepts, PICNET is different from these previous projects, as it works from the opposite direction and collects images that are representative for a given concept. Hence, an ideal image in PICNET will represent only one salient concept, rather than a multitude of concepts as in the ESP or the Google Image Quiz games. Once again, in the terminology used in this paper, a concept is defined as "an abstract idea or mental symbol, inferred or derived from specific instances." Thus, our goal in PICNET is to find concept-image associations so that a concept can be inferred, as unambiguously as possible, from its correspond-

[1] http://www.gamesforthebrain.com/game/imagequiz/

ing image. As with the concept-word associations that are used in language, these are only approximations of the meaning of a concept.

3. PicNet

PicNet is a knowledge-base consisting of dual visual-linguistic representations for words and phrases – seen as the smallest units of communication that carry meaning. Starting with a machine readable dictionary that defines the words in the common vocabulary and their possible meanings, PICNET seeks to add visual representations to the dictionary entries, to the end of building a knowledge-base that combines both verbal and visual representations of these basic concepts.

In addition to the role played in the generation of "pictorial representations," as described in Section 4, PICNET can also be used in other applications, or as a standalone resource. First, PICNET can be seen as an associative facilitator for children learning how to read or for people learning a second language. Research has shown that combining pictures with text can improve the learning process for children or second language learners (Glenberg *et al.*, 2004). Moreover, studies have shown that children from different countries, not speaking each other's language, were able to communicate about children stories just by using drawings and pictures (Alburo *et al.*, 2005). While understanding the word *apple* requires knowledge of English, understanding a picture that represents an *apple* is transparent to languages – and such a representation can be understood by speakers of any language, regardless of their origins or literacy.[2]

Second, by exploiting the language-independent aspect of visual representations, PICNET could provide the means for building multi-lingual dictionaries and semantic knowledge-bases, without requiring bilingual speakers. For instance, by showing an image associated with a given concept and asking users to provide the corresponding words in the various languages, can result into a self-managed multi-lingual extension mechanism for knowledge-bases like WordNet.

Third, PICNET has also the potential to bridge the gap between research work in image processing and language processing. For instance, the explicit concept/image associations available in PICNET, as well as the hierarchical relations among pictures automatically derived from the corresponding semantic network (WordNet) can help improve the quality of systems for image retrieval and/or classification. Similarly, image content analysis can help lan-

[2] Note that PICNET does not limit the number of images that can be associated with a given concept. It is not expected that one image can always fully represent a particular concept, but a set of images taken together may represent the collective notion of diverse contributors, and provide an expanded understanding to a user of the system.

guage processing tasks. For example, (Barnard *et al.*, 2003) showed how text and image information can be jointly used to improve the performance of a word sense disambiguation task. Other tasks, such as information extraction, information retrieval, named entity recognition, etc., could be also improved using dual visual/linguistic representations.

3.1. Constructing PicNet

PicNet relies on a Web-based system for augmenting dictionaries with illustrative images using volunteer contributions over the Web. The assumption is that all Web users are experts when it comes to understanding the content of images and finding associations between words and pictures. Given a word and its possible meanings – as defined by a comprehensive dictionary – Web users participate in a variety of game-like activities targeting the association of pictures with words. The first iteration of PicNet focuses on nouns and verbs, with an emphasis on concrete nouns and noun-phrases (Coltheart, 1981). No attempt is made at this point to assign pictures to adjectives and adverbs.

The primary lexical resource used in PicNet is WordNet (Miller, 1995) – a machine readable dictionary containing a large number of concepts and relations between them. The original WordNet dictionary covers English concepts, and it is also linked to a large number of dictionaries covering several European languages (Vossen, 1998; Tufis and Cristea, 2002), as well as to the Chinese HowNet dictionary (Carpuat *et al.*, 2002).

Initially, PicNet was seeded with images culled from automated image searches using PicSearch[3] and AltaVista,[4] which resulted in 72,968 word/image associations automatically collected. The results of the automatic process were particularly good when searching for concrete nouns or specific entities with precise definitions, such as a particular plant genus. However, in general, the value of the automated search results is mixed. It seems that the search engines' procedure relies more on the image filename rather than the textual context of the image. Also, due to the sheer quantity of synsets given and the search and processing time required, no particular attempt was made to differentiate between different senses of a word when performing the automated seeding.

3.2. Activities in PicNet

The validation of the entries in PicNet is performed by Web volunteers who can choose to participate in a variety of activities, including free association (assign a concept to a randomly selected image from the database), image

[3] http://www.picsearch.com
[4] http://www.altavista.com/image

upload (upload an image the user finds representative for a given concept), image validation (assign a quality vote to a randomly selected concept/image association from the PICNET dictionary), or competitive free association (a game-like activity where multiple users can simultaneously vote on a concept/image association). When a user logs on the PICNET website, she can choose to participate in any of the following activities:

Upload Images. Given a word with a meaning, as defined by the dictionary, a user may upload an image that she finds representative for the given concept. Images contributed in this manner are not immediately associated with a concept, but will gain their word associations in other PICNET phases.

Free Association. In this task, the user is shown a random image from the dictionary. The image may or may not already have a concept association, which is not apparent to the user in order to avoid bias. The user is asked to enter a concept related to the image, and a new concept/image association is created.

Validate Images. Concept/image associations are created by user uploads, user free association, or the initial automated PICNET seeding. In this task, users are shown a concept/image association randomly drawn from the PICNET dictionary. The user is shown the concept with its dictionary definition and the associated image, and she may then vote on the appropriateness of the image. Figure 1 shows a snapshot of the validation screen.

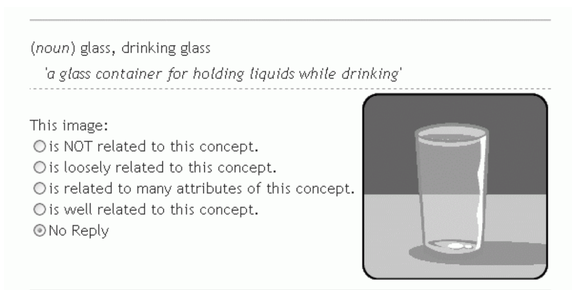


Figure 1. Validation screen in PICNET.

Competitive Free Association. To help motivate users to participate, a game process was also implemented to allow users to compete with each other. The game begins once a minimum of five players join and a majority votes to start. To start the round, each player is shown an image from the PICNET database and asked to provide an anonymous word association. Identical entries from multiple players are coalesced. After all players have entered a suggestion, the players vote for the best selection which is not their own. The player who entered the concept receiving the most votes wins the round. If multiple players entered the winning concept, the points for the round are split. The

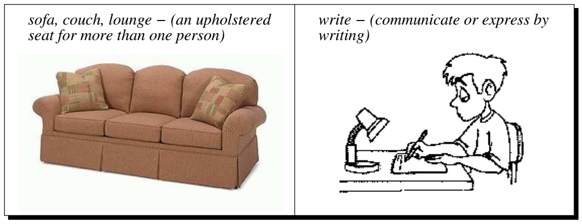


Figure 2. Sample concept/image associations from PICNET.

concept/image associations are added to the PICNET database, with the score value equal to the number of votes received.

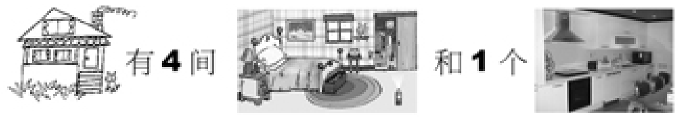
3.3. ENSURING DATA QUALITY

Collecting from the general public holds the promise of providing much data at low cost. It also makes attending to an important aspect of data collection: ensuring contribution quality. PICNET implements a scoring scheme that ranks concept/image pairs based on the total number of votes received from users of the various PICNET activities. A complete history of users' decisions is maintained and used to rank the concept/image associations. Each action provides an implicit quantified vote relating to the concept/image pair. The sum of these votes creates a score for the pair, allowing PICNET to rank images associated to a particular concept. The following list represents the possible actions that users can perform on the PICNET site, and the corresponding votes: Upload an image for a selected concept (+5); Image validation – well related to the concept (+4); Image validation – related to many concept attributes (+3); Image validation – loosely related to the concept (+1); Image validation – not related to the concept (-5); Free association (+3); Competitive free association (+n, where n is the number of users agreeing with the association).

3.4. PICNET EVALUATIONS

Evaluations concerning the quality of the data collected through PICNET were conducted based on the concept/image associations collected up-to-date for approximately 6,200 concepts from 320 contributors.

First, for the free association activity, the average concurrence among users voting on the appropriateness of an image was measured, resulting in an average concurrence of 43% with a standard variance of 0.05. Note that this is "free" association, and thus for an image representing e.g., an *African violet*, several labels are possible including *flower*, *violet*, *African violet*. Only perfect label matches are counted toward the agreement measure, and thus the concurrence of 43% indicates a consistent agreement.



(a) Pictorial representation for *"The house has four bedrooms and one kitchen."*

(b) Mixed pictorial and linguistic representation (automatic) for *"You should read this book.'*

I eat the egg and the coffee work as breakfast.

(c) Linguistic representation (automatic) for *"I eat eggs and coffee for breakfast."*

Figure 3. Sample pictorial and linguistic representations for three input texts.

Second, the quality of the concept/image associations was evaluated by a trusted human judge. For each association from a randomly selected set of 200 concept/image associations, the human judge assigned one of the following four options: closely related, related, loosely related, not related. Table I shows the results of the evaluation. Overall, 81% of the images were found to be related to their corresponding words. Only 11.5% of the concept/image associations were incorrect, which is remarkable, given that the sampling of the evaluation data set was random.

Table I. Evaluation of concept/image associations in PICNET.

Rank	Count	Percentage
Closely related	110	55.0%
Related	52	26.0%
Loosely related	15	7.5%
Not related	23	11.5%

Figure 2 shows two sample concept/image associations collected with PICNET and their dictionary definitions.

4. Understanding with Pictures

Starting with PICNET, we implemented a system for the automatic construction of pictorial representations for simple sentences. The hypothesis guiding our experiments is that simple sentences can be conveyed via pictorial representations with limited or no use of linguistic descriptions. While linguistic expressions are certainly preferred when it comes to complex, abstract concepts such as *materialism* or *scholastics*, simple concrete concepts such as *apple* or *drink* can be effectively described through pictures, and consequently create pictorial representations of information.

Our goal is to test the level of understanding for *entire* pieces of information represented with pictures, e.g. short sentences such as *I want to drink a glass of water*, which is different than testing the ability to grasp a single concept represented in a picture (e.g. understand that the concept shown in a picture is *apple*). We therefore perform our experiments within a translation framework, where we attempt to determine and evaluate the amount of information that can be conveyed through pictorial representations.

Specifically, we compare the level of understanding for three different ways of representing information: (1) fully conceptual, using only pictorial representations; (2) mixed linguistic and conceptual, using representations consisting of pictures placed within a linguistic context; and finally (3) fully linguistic, using only words to represent information.

4.1. A SYSTEM FOR THE CONSTRUCTION OF PICTORIAL REPRESENTATIONS OF SIMPLE SENTENCES

Starting with an input sentence, the text is tokenized and part-of-speech tagged (Brill, 1992), and word lemmas are identified using a WordNet-based lemmatizer. Next, we attempt to identify the most likely meaning for each open-class word using a publicly available state-of-the-art sense tagger that identifies the meaning of words in unrestricted text with respect to the WordNet sense inventory (Mihalcea and Csomai, 2005).

Once the text is pre-processed, and the open-class words are labeled with their parts-of-speech and corresponding word meanings, we use PICNET to identify pictorial representations for each noun and verb. We supply PICNET with the lemma, part-of-speech, and sense number, and retrieve the highest ranked picture from the collection of concept/image associations available in PICNET. To avoid introducing errors in the pictorial representation, we use only those concept/image associations that rank above a threshold score of 4, indicating a high quality association.

Once again, pictorial representations are assigned only to nouns and verbs, and no attempt is made to assign pictures to adjectives or adverbs. In addition to the image representations for nouns and verbs as collected through PICNET, we also use a set of pictorial representations for pronouns, using images from a language learning course.[5]

[5] http://tell.fll.purdue.edu/JapanProj/FLClipart/

4.2. Representation Scenarios

We conduct out experiments under the assumption that there is a language barrier between the two participants in an information communication process. The sender (speaker) attempts to communicate with a receiver (listener), but the only communication means available is a language known to the sender, but not to the receiver. We therefore deal with a standard translation framework, where the goal is to convey information represented in an "unknown" (source) language to a speaker of a "known" (target) language. The following three translation scenarios are evaluated:

Scenario S1. No language translation tool is available. The information is conveyed exclusively through pictures, and while linguistic representations can still be used to suggest the presence of additional concepts, they are not understood by the information recipient. In this scenario, the communication is performed entirely at conceptual level. Figure 3(a) shows an example of such a pictorial representation.

Scenario S2. An automatic language translation tool is available, which is coupled with a tool for constructing pictorial representations, for a dual visual-linguistic representation. The linguistic representations in the target ("known") language are produced using an automatic translation system, and therefore not necessarily accurate. Figure 3(b) shows an example of a mixed pictorial-linguistic representation.[6]

Scenario S3. The third case we evaluate consists of a standard language translation scenario, where the information is conveyed entirely at linguistic level. Similar with the previous case, the assumption is that a machine translation tool is available, which can produce (sometime erroneous) linguistic representations in the target "known" language. Unlike the previous scenario however, no pictorial representations are used, and therefore we evaluate the understanding of information using representations that are fully linguistic. An example of such a representation is illustrated in Figure 3(c).

5. Evaluations

In order to evaluate the quality of the pictorial representations, we created a testbed of 50 short sentences, consisting of 30 randomly selected examples from language learning courses, and 20 sentences from various domain-specific texts covering fields such as finance, sports, travel, etc. While all the sentences in our testbed are short, with an average of 15 words each, they have various levels of difficulty, ranging from simple basic vocabulary

[6] The English words "this" and "should" were obtained by automatically translating the Chinese words for which no pictorial representations were available.

taught in beginner language classes, to more complex sentences containing domain-specific vocabulary.

Although our system, as described in Section 4.1, is designed to work with English as a source language, in order to facilitate the evaluations we have also created a Chinese version of the sentences in our data set.[7] The reason for using Chinese (rather than English) as the source "unknown" language was to ensure the fairness of the evaluation: since this research was carried out in an English-speaking country, it was difficult to find users who did not speak English and who were completely unaware of the peculiarities of the English language. Instead, by using Chinese as the source language, we were able to conduct an evaluation where the users interpreting the pictorial representations were not aware of any of the specifics of the source language.

One possible criticism of this experimental design is that the Chinese translations may be more "English like" in terms of structure than if we were to start with spontaneous Chinese text. While we made our best attempt to generate natural and correct Chinese sentences by relying on two native Chinese speakers, if it is still true that the Chinese text follows the English structure more than an original Chinese text, the same is true for all the three scenarios below, and thus the comparative evaluations embed the same potential benefit of an "easier" source language across all our experiments.

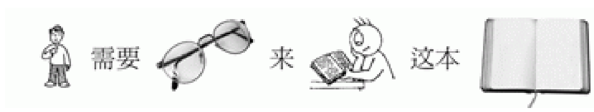
For each sentence in our data set, three representations were generated:

1. A **pictorial representation**, where verbs, nouns, and pronouns are represented with pictures, while the remaining context is represented in Chinese. The pictorial representations, automatically generated for the English version of each sentence, were manually assigned to the concepts in the Chinese sentence. It is important to note that this step was required exclusively for the purpose of conducting the evaluations. In the general case, the pictorial representations are automatically assigned to a source English sentence, and used as such in the communication process. However, since we wanted to circumvent the problem of all the users available for our study being English speakers, we chose to conduct the evaluations using a language different than English (and consequently selected Chinese as the source language). Note that no pictorial representations are generated for those verbs or nouns not available in PICNET.

2. A **mixed pictorial and linguistic representation**, where verbs, nouns, and pronouns are still represented with pictures, but the context is represented in English.

[7] This represents the "unknown" language for the translation evaluations below. The translation was generated by two native Chinese speakers, who took care to produce coherent Chinese sentences that were adequate as well. When there were disagreements, there was further discussion to come up with a better translation agreed by the two.

3. A **linguistic representation**, as obtained from a machine translation system (Systran http://www.systransoft.com), which automatically translates the Chinese version of each sentence into English. No pictorial representations are used in this scenario. Note that the performance obtained with Systran is higher than the one that would be obtained with a word-by-word dictionary-based approach, which depends only on bilingual dictionaries. The quality of the linguistic translation is thus higher than what could be obtained using only a dictionary (as we do in our PICNET-based translations).



Interpretation 1:	*I use glasses to read my books.*
Interpretation 2:	*I need glasses to read a book.*
Interpretation 3:	*I need my eye glasses to read this book.*

Figure 4. Various interpretations by different users for a sample pictorial representation.

Each of the three representations was then shown to fifteen different users, who were asked to indicate in their own words their interpretation of the visual and/or linguistic representations. The users were instructed to write a complete, grammatically-correct sentence that "tells a story," i.e., a sentence that closely matches the set of pictures provided. The users were told that each picture may substitute for a noun, pronoun or verb, that the pictures flow in a left-to-right sequence, and that the order of the pictures is not necessarily reflecting the word order in the final sentence.

Figure 4 shows a pictorial representation for the sentence *"I need glasses to read this book,"* and three interpretations by three different users.[8]

5.1. EVALUATION METRICS

To assess the quality of the interpretations generated for each of the three representation scenarios described before, we use both manual and automatic assessments of quality, based on metrics typically used in machine translation evaluations.

First, we use a human evaluation of quality, consisting of an adequacy assessment. A human judge was presented with the correct English reference and a candidate interpretation, and was asked to indicate how much of the information in the gold standard reference sentence was preserved in the

[8] A pictorial representation was not used for the verb *"need"*, since no image association was found in PICNET for this concept.

candidate interpretation. The assessment is done on a scale from 1 (“none of it”) to 5 (“all the information”).[9]

Second, we use two automatic evaluations of quality traditionally used in machine translation evaluation. The NIST evaluation (Doddington, 2002) is based on the Bleu score (Papineni *et al.*, 2002). It is an information-weighted measure of the precision of unigrams, bigrams, trigrams, four-grams, and five-grams in the candidate interpretations with respect to the “gold-standard” reference translation. The other metric is the GTM score (Turian *et al.*, 2003), which measures the similarity between texts in terms of precision, recall, and F-measure. Both measures were found to have good performance at discriminating translation quality, with high correlations to human judgments.

A possible criticism of our evaluations is the fact that the pictorial representations are not generated by a full end-to-end system that starts with the Chinese source sentences, but rather we use the English correspondent sentences to find the pictorial representations in PICNET.[10] It is important however to note that the goal of these evaluations is to compare the **interpretation** that humans can produce based on different “foreign languages” where, for the purpose of our experiments, the pictorial representation is regarded as a foreign language as well. We compare the interpretation of the pictorial representations with the interpretation of the Chinese sentences. The English translation of the Chinese sentences is an attempt to simulate the benefit that one would have from using machine translation software, and thus the translation itself can be regarded as a “foreign language” for which we measure its interpretability. Hence, one should look at these experiments and evaluations as a comparison of the **interpretation** that humans produce for different representations, and not a comparison of the representations themselves or the mechanisms used to generate such representations.

5.2. RESULTS

For each sentence in our testbed and for each possible visual or linguistic representation, we collected interpretations from fifteen different users, accounting for a total of 2,250 interpretations. No Chinese speakers were allowed to participate in the evaluations, since Chinese was the “unknown” language used in our experiments. The user group included different ethnic

[9] Traditionally, human evaluations of machine translation quality have also considered fluency as an evaluation criterion. However, since we measure the quality of the *human*-produced *interpretations* (rather than measuring directly the quality of the automatically produced translations), the interpretations are fluent, and therefore do not require an explicit evaluation of fluency.

[10] As explained before, the reason for this convoluted design was not simply to work with a language that was easier to analyze, but more importantly, to be able to use the resources that are not quite as extensive in other languages (e.g., there is no Chinese PicNet, and the Chinese WordNet is not as well developed as the English WordNet).

groups, e.g. Hispanics, Caucasians, Latin Americans, Indians, accounting for different cultural biases. While all the users were accustomed to the American culture (all of them having lived in the United States for two or more years), only a small fraction of them were English native speakers.

All the interpretations provided by the users were scored using the three evaluation measures: the GTM F-measure and the NIST scores, and the manually assessed adequacy. Table II shows the evaluation results, averaged across all users and all sentences.

Table II. Results for the three representation scenarios, using automatic and manual evaluation criteria. Standard deviations were measured at: 7.80 for the NIST score, 6.30 for the GTM score, and 0.31 for the adequacy score.

	Evaluation		
	automatic		manual
Type of translation	NIST (Bleu)	GTM	Adequacy
S1: Pictures	41.21	32.56	3.81
S2: Pictures+linguistic	52.97	41.65	4.32
S3: Linguistic	55.97	44.67	4.40

The lower bound is represented by the "no communication" scenario (no language-based communication between the two speakers), corresponding to a baseline score of 0 for all the translation scores. For the human adequacy score, the upper bound consists of a score of 5, which reflects a perfect interpretation. For the NIST and the GTM scores, it is difficult to approximate an upper bound, since these automatic evaluations do not have the ability to account for paraphrases or other semantic variations, which typically get penalized in these scores. Previous evaluations of a NIST-like score on human-labeled paraphrases led to a score of 70%, which can be considered as a rough estimation of the upper bound (Mihalcea *et al.*, 2006).

6. Discussion and Data Analysis

The results indicate that a significant fraction of the information contained in simple sentences can be conveyed through pictorial representations. The human adequacy score of 3.81, also reflected in the automatic NIST and GTM scores, indicate that about 76%[11] of the content can be effectively

[11] The fraction of the adequacy score for pictorial representations (3.81) divided by the maximum adequacy score (5.00).

communicated using pictures. This score is explained by the intuitive visual descriptions that can be assigned to some of the concepts in a text, and by the humans ability to efficiently *contextualize* concepts using their background world knowledge. For instance, while the concepts *read* and *book* could also lead to a statement such as e.g. "*Read about a book*," the most likely interpretation is "*Read a book*," which is what most people will think of when seeing the pictorial representations of these two concepts.

6.1. Data Analysis

In an attempt to understand the level of difficulty associated with the understanding of pictorial representations for different sentence types, we performed a detailed analysis of the test set, and measured the correlation between various characteristics of the test sentences and the level of understanding achieved during the sentence interpretation experiments. Specifically, given a sentence feature (e.g. the number of words in a sentence), and an evaluation score for interpretation quality (e.g. the NIST score), we determined the Pearson correlation factor (r) between the feature considered and the quality of the interpretation. In all the correlation experiments, we report correlation measures using the NIST evaluation scores, but similar correlation scores were observed for the other evaluation metrics. As typically assumed in previous correlation studies, a Pearson factor of $0.10 - 0.29$ is associated with a *low* correlation, $0.30 - 0.59$ represents a *medium* correlation, and $0.60 - 1.00$ is considered *high* correlation.

Based on correlation analyzes for a number of features, the following observations were drawn.

Sentence length. There is a high negative correlation ($r = -0.67$) between the number of words in a sentence and the level of understanding achieved for the pictorial representations. This suggests that the understanding of pictorial representations increases with decreasing sentence length. Our pictorial representation paradigm is therefore most effective for short sentences.

Ratio of words with a given part-of-speech. There is a medium positive correlation ($r = 0.44$) between the proportion of nouns in a sentence and the level of understanding, and a medium negative correlation ($r = -0.47$) between the number of function words and the quality of interpretation, indicating that sentences that are "dense" in concepts (large number of nouns, small number of function words) are easier to understand when represented through pictures.

Syntactic complexity. We modeled syntactic complexity by counting the number of different syntactic phrases (e.g. noun phrases), and by determining the high-level structure of the syntactic parse tree (e.g. subject-verb, subject-verb-indirect_object, etc.). We found that the understanding of pictorial representations decreases with increasing syntactic complexity, with a medium negative correlation observed between the number of noun-phrases ($r = -0.49$) or prepositional phrases ($r = -0.51$) in a sentence and the quality of interpretation. Although no significant correlation was found between the level of understanding of a pictorial representation and the structure of the syntactic parse tree, on average better interpretations were observed for sentences with a complete subject-verb-direct_object structure (as compared to e.g. sentences with a subject-verb structure).

Semantic classes. Using the semantic classes from WordNet (26 semantic classes for nouns and 15 semantic classes for verbs), we determined for each sentence the number of concepts belonging to each semantic class, and measured the correlation with the level of understanding for pictorial representations. We found a low positive correlation ($r = 0.20 - 0.30$) associated with the number of nouns belonging to the semantic class "animal" (e.g. *dog*) and "communication" (*e.g. letter*) and the verbs from the semantic classes of "cognition" (e.g. *read*) and "consumption" (e.g. *drink*). No significant correlations were found for the other semantic classes.

Word frequency. For each of the sentences in the test set, we determined the frequency of each constituent word (excluding stopwords) using the British National Corpus. These word frequencies were then combined into a score which, after normalization with the length of the sentence, reflects the usage frequency for the concepts described in a sentence. We found a medium positive correlation ($r = 0.38$) between the combined frequency of the words in a sentence and the level of understanding for pictorial representations, suggesting that it is easier to understand and interpret the pictorial descriptions associated with frequently used words.

6.2. Score Analysis

An analysis of the scores listed in Table II reveals interesting aspects concerning the amount of understanding achieved for different scenarios.

The score achieved through the pictorial representations alone (S1) represents a large improvement over the score of 0 for the "no communication" baseline (which occurs when there are no means of communication between the speakers). The score achieved by this scenario indicates the role played by conceptual representations (pictures) in the overall understanding of simple sentences.

The difference between the scores achieved with scenario S1 (pictorial representations) and scenario S2 (mixed pictorial and linguistic representations) points out the role played by context that cannot be described with visual representations. Adjectives, adverbs, prepositions, abstract nouns and verbs, and others constitute a linguistic context that cannot be represented with pictures, and which nonetheless has an important role in the communication process. While simple syntactic structures as reflected by word order can be represented in the pictorial interpretations (e.g., subject-verb-object structures), more complex linguistic structures such as coreference, embedded clauses, etc., are lacking in the visual representations, contributing to possible interpretation errors.

Finally, the gap between the second (S2) and the third (S3) scenarios indicates the advantage of words over pictures for producing accurate interpretations. Note however that this is a rather small gap, which suggests that pictorial representations placed in a linguistic context are intuitive, and can successfully convey information across speakers, with an effectiveness that is comparable to full linguistic representations.

There were also cases when the pictorial representations failed to convey the desired meaning. For instance, the illustration of the pronoun *he*, a *riverbank*, and a *torch* (for *He sees the riverbank illuminated by a torch*) received a wrong interpretation from most users, perhaps due to the unusual, not necessarily commonsensical association between the *riverbank* and the *torch*, which most likely hindered the users ability to effectively contextualize the information.

Interestingly, there were also cases where the interpretation of the pictorial representation was better than the one for the linguistic translation. For instance, the Chinese sentence for *I read email on my computer* was wrongly translated by the machine translation system to *I read electricity on my computer post.* which was misleading, and led to an interpretation that was worse than the one generated by the illustration of the concepts of *I*, *read*, *email*, and *computer*.

Overall, while the interpretation of visual representations is subject to several limitations, the understanding achieved based on pictorial representations for simple short sentences was found to be within a comparable range of the understanding achieved based on an automatic machine translation system, which suggests that such pictorial representations can be used for the purpose of communicating simple pieces of information.

7. Conclusions

In this paper, we described the construction of an illustrated dictionary, which brings together in one resource the linguistic and visual representations of words and phrases. We proposed “pictorial representations” as a means for conveying simple pieces of information across language barriers, and showed how our our illustrated dictionary can be used to build a system that can generate pictorial representations for simple sentences. Comparative experiments conducted on visual and linguistic representations of information have shown that a considerable amount of understanding can be achieved through pictorial descriptions, with results within a comparable range of those obtained with current machine translation techniques.

Future work will consider the analysis of more complex sentences of various degrees of difficulty, as well as input with alternate word orders and different linguistic structure. We also plan to experiment with modifiers such as color coding, relative positions, size of pictures in a sequence, to highlight the different aspects of a pictorial representation. Cultural differences in picture interpretation are also an interesting aspect that we plan to consider in future evaluations, through evaluations with subjects from diverse cultural backgrounds (e.g., subjects whose native language does not have a subject-verb-object order).

Acknowledgments

The authors are grateful to the journal editors and to the anonymous reviewers for their useful comments, which helped improve an earlier version of this paper.

Appendix

The following sentences were used in the evaluations:

This cat is four years old.
He gives the child a dime.
You can buy a used car at a low cost.
Please sit down on this chair.
I am taking a computer course at a local college.
You should go to a doctor for that cold.
Cotton is used to make clothes.
I read his latest column in the New York Times.
I eat an apple after dinner.
The bank closes at three in the afternoon.
He bought a new boat for his birthday.
Can you get some bread from the supermarket?
My brother lives in Seattle.
You should read this book.
Will you like to go dancing with me this Saturday?
I visited my dad last week.
Will you like the boiled egg or fried egg?
He milks the cow everyday.
He drinks two glasses of water.

I eat eggs and coffee for breakfast.
I will travel to Africa.
I bought a pair of new shoes last week.
There have been three tornadoes in Oklahoma.
I need my glasses to read this book.
I wrote a letter to my mother.
I read email on my computer.
Please bring me a glass of tea and a biscuit.
The house has four bedrooms and one kitchen.
I go to the gym to exercise.
I like to eat milk and cereal in the morning.
He saw the sign above the door of the hut: Home Sweet Home.
He dumped the pan of crumbled hardtack into the boiling pot of lobscouse.
He settled on the sofa with his coffee, warming his hands on the cup.
He found the pilot light and turned on one of the burners for her.
The portable record player with a pile of classical records beside it.
He reached out and felt the bath towel hanging on the towel rack over the tub.
They took Jesus's body, then, and wrapped it in winding-clothes with the spices.
David reached for the pair of pistols in the saddlebags at his feet.
The fish took the bait.
He could see the bright torches lighting the riverbank.
In the corner was the soldier with the white flag.
She lay still on the bed, her head hardly denting the pillow.
Her legs hung down long and thin as she sat on the high stool.
He finally fell asleep around six in the morning with the aid of a sleeping pill.
In one hand she clutched a hundred dollar bill and in the other a straw suitcase.
That couple has a son and a daughter.
Tanks lined up at the border will be no more helpful.
The sick were always receiving medicines.
The bottle was filled with flour.
There was a lady there, in pyjamas.

References

J. Alburo, A. Komlodi, J. Preece, A. Druin, A. Elkiss, and P. Resnik. Evaluating a cross-cultural children's online book community: Sociability, usability, and cultural exchange. Technical report, University of Maryland, 2005.

N. Alm, M. Iwabuchi, P. Andreasen, and K. Nakamura. A multi-lingual augmentative communication system. In *Univeral Access: Theoretical Perspectives, Practice and Experience, Springer*, pages 398–408, Paris, France, 2002.

K. Barnard and D.A. Forsyth. Learning the semantics of words and pictures. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 408–415, Vancouver, Canada, 2001.

K. Barnard, M. Johnson, and D. Forsyth. Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data*, Edmonton, Canada, 2003.

M. Behrmann and S. Byng. *Cognitive Neuropsychology in Clinical Practice*. Oxford University Press, 1992.

K. Bloomberg, G. Karlan, and L. Lloyd. The comparative translucency of initial lexical items represented in five graphic symbol systems and sets. *Journal of Speech and Hearing Research*, 33(4):717–725, 1990.

M. Boshernitsan and M. Downes. Visual programming languages: A survey. Technical report, U.C. Berkeley, 1997.

E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, Italy, 1992.

R. Carney and J. Levin. Pictorial illustration still improve students' learning from text. *Educational Psychology Review*, 14(1):5–26, 2002.

M. Carpuat, G. Ngai, P. Fung, and K. Church. Creating a bilingual ontology: A corpus-based approach for aligning WordNet and HowNet. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 284–292, Taipei, Taiwan, 2002.

S. Chang, G. Costagliola, S. Orefice, G. Polese, and B. Baker. A methodology and interactive environment for iconic language design with applications to augmentative communication. In *Proceedings of the IEEE workshop on visual languages*, pages 110–116, Seattle, WA, 1992.

S. Clay and J. Wilhelms. Put: Language-based interactive manipulation of objects. *IEEE Computer Graphics and Applications*, 16(2):31–39, 1996.

M. Coltheart. The mrc psycholinguistic database. *Quarterly Journal of Experimental Psychology*, 33(4):497–505, 1981. available online at http://www.psy.uwa.edu.au/MRCDataBase/uwa_mrc.htm.

B. Coyne and R. Sproat. Wordseye: An automatic text-to-scene conversion system. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques*, pages 487–496, Los Angeles, CA, 2001.

G. Doddington. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of Human Language Technology HLT-2002*, pages 138–145, San Diego, 2002.

2005. http://www.ethnologue.com.

D. Evans, L. Bowick, M. Johnson, and P. Blenkhorn. Using iconicity to evaluate symbol use. In *Proceedings of the 10th International Conference on Computers Helping People*, pages 874–881, Linz, Austria, 2006.

W.W. Gibbs. Saving dying languages. *Scientific American*, 287(2):79–86, 2002.

A. M. Glenberg, T. Gutierrez, J. R. Levin, S. Japuntich, and M. P. Kaschak. Activity and imagined activity can enhance young children's reading comprehension. *Journal of Educational Psychology*, 96(3):424–436, 2004.

E. Hanson and A. Hartzema. Evaluating pictograms as an aid for counselling elderly and low-literate patients. *Journal of Pharmaceutical Marketing Management*, 9(3):41–54, 1995.

L. Haupt and E. Alant. The iconicity of picture communication symbols for rural Zulu children. *South African Journal of Commun Disorders*, 49:40–49, 2003.

M. Huer. Examining perceptions of graphic symbols across cultures: Preliminary study of the impact of culture/ethnicity. *Augmentative and Alternative Communication*, 16(3):180–185, 2000.

R. Johansson, A. Berglund, M. Danielsson, and P. Nugues. Automatic text-to-scene conversion in the traffic accident domain. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1073–1078, Edinburgh, Scotland, 2005.

I. Medhi, A. Sagar, and K. Toyama. Text free user interfaces for iliterate and semi literate users. In *International conference on information and communication technologies and development*, pages 72–82, Berkeley, CA, 2006.

R. Mihalcea and A. Csomai. Senselearner: Word sense disambiguation for all words in unrestricted text. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics*, pages 53–56, Ann Arbor, MI, 2005.

R. Mihalcea, C. Corley, and C. Strapparava. Corpus-based and knowledge-based approaches to text semantic similarity. In *Proceedings of the American Association for Artificial Intelligence*, pages 775–780, Boston, MA, 2006.

G. Miller. Wordnet: A lexical database. *Communication of the ACM*, 38(11):39–41, 1995.

M. Mizuko. Transparency and ease of learning of symbols represented by blissymbols, pcs and picsyms. *Augmentative and Alternative Communication*, 3(3):129–136, 1987.

C. Musselwhite and D. Ruscello. Transparency of three communication symbol systems. *Journal of Speech and Hearing Research*, 27(3):436–443, 1984.

K. Nakamura, A. Newell, N. Alm, and A. Waller. How do members of different language communities compose sentences with a picture-based communication system? – a cross cultural study of picture-based sentences constructed by english and japanese speakers. *Augmentative and Alternative Communication*, 14(2):71–79, 1998.

R. Nigam. Do individuals from diverse cultural and ethnic backgrounds perceive graphic symbols differently? *Augmentative and Alternative Communication*, 19(2):135–136, 2003.

J.Y. Pan, H.J. Yang, C. Faloutsos, and P. Duygulu. Gcap: Graph-based automatic image captioning. In *Proceedings of the 4th International Workshop on Multimedia Data and Document Engineering*, pages 146–155, Washington, DC, 2004.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, 2002.

M.C. Potter, J.F. Kroll, B. Yachzel, E. Carpenter, and J. Sherman. Pictures in sentences: understanding without words. *Journal of Experimental Psychology*, 115(3):281–294, 1986.

2008. http://sharing.mayer-johnson.com/Portals/0/NTForums_Attach/Symbolate.pdf.

D. Tufis and D. Cristea. Methodological issues in building the Romanian WordNet and consistency checks in Balkanet. In *Proceedings of the LREC Workshop on Wordnet Structures and Standardisation*, pages 35–41, Las Palmas, Spain, 2002.

J. Turian, L. Shen, and I. D. Melamed. Evaluation of machine translation and its evaluation. In *Proceedings of the Machine Translation Summit*, pages 386–393, New Orleans, LA, 2003.

L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proceedings of the Conference on Human factors in computing systems*, pages 319–326, Vienna, Austria, 2004.

P. Vossen. *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Kluwer Academic Publishers, Dordrecht, 1998.

A. Yamada, T. Yamamoto, H. Ikeda, T. Nishida, and S. Doshita. Reconstructing spatial image from natural language texts. In *Proceedings of the 14th International Conference on Computational linguistics*, pages 1279–1283, 1992.

X. Zhu, A. B Goldberg, M. Eldawy, C. R. Dyer, and B. Strock. A text-to-picture synthesis system for augmenting communication. In *Proceedings of the American Association for Artificial Intelligence: Integrated Intelligence Track*, pages 1590–1595, Vancouver, Canada, 2007.