TOWARDS COMMUNICATING SIMPLE SENTENCES USING PICTORIAL REPRESENTATIONS

Chee Wee Leong, B. Eng.

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

May 2006

APPROVED:

Rada Mihalcea, Major Professor
Paul Tarau, Committee Member
Elizabeth Figa, Committee Member
Armin Mikler, Graduate Advisor of Computer Science and Engineering
Krishna Kavi, Chair of the Department of Computer Science and Engineering
Oscar Garcia, Dean of the College of Engineering
Sandra L. Terrell, Dean of the Robert B. Toulouse School of Graduate Studies

Leong, Chee Wee, Towards Communicating Simple Sentence using Pictorial Representations. Master of Science (Computer Science), May 2006, 64 pp., 3 tables, 11 illustrations, bibliography, 26 titles.

Language can sometimes be an impediment in communication. Whether we are talking about people who speak different languages, students who are learning a new language, or people with language disorders, the understanding of linguistic representations in a given language requires a certain amount of knowledge that not everybody has. In this thesis, we propose "translation through pictures" as a means for conveying simple pieces of information across language barriers, and describe a system that can automatically generate pictorial representations for simple sentences. Comparative experiments conducted on visual and linguistic representations of information show 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. Moreover, a user study conducted around the pictorial translation system reveals that users found the system to generally produce correct word/image associations, and rate the system as interactive and intelligent.

ACKNOWLEDGMENTS

I am indebted to my research supervisor, Dr. Rada Mihalcea, for her dedication to this project. It's she who started me on the great idea of doing research in natural language processing. Her contribution to this thesis is immense, and certainly crucial, to its success. The class times, research meetings, brainstorming sessions are simply the best of my academic endeavors to date.

I also want to convey my heartfelt thanks to Dr. Paul Tarau and Dr. Elizabeth Figa, who took the invitation and stood on the examination committee. Their constructive feedback had certainly improved the overall quality of this thesis.

To the members of the language information technology group - especially Andras, Chris and Samer. Thank you all for the support on this project. I appreciate your help in giving evaluations, the critical, the better. Thanks for taking time out of nowhere to do the survey, and of course, debugging the programs when i needed a helping hand.

For my fiancée, Chinvy. Special gratitude to you for the emotional support and prayers throughout. Surely, I won't have made it without your persistent encouragement.

I want to hug my parents the next time I would see them, who took the pains to send me to United States for further education. Also, to my sister, Amy, and brother-in-law, Holy, for their faithful support.

To Karl and Leon, my mentors here in Texas, for believing in my ability to go all the way.

Most of all, for the Creator and Maker of the Universe, the Lord our God. Praise be to You who is the source of all knowledge and wisdom.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Natural languages are formed for the purposes of communicating daily experiences, conveying thoughts and emotions, educating ideologies, and others. Essentially, any natural language can be described by a system of symbols (lexemes) and the grammars (rules) which manipulate the symbols. By and large, it is the symbols, grammars, and their interactions that make a particular language unique from others, although certain languages do have overlapping sets of symbols or grammars. For instance, “cognates” represent words that have similar spelling across languages, usually identified based on some string similarity measure (like longest common subsequence, or Levinstein distance). Such a string similarity will indicate that e.g. “name” (English) and “nome” (Italian) and “nume” (Romanian) refer to the same concept, and were derived from the same root. Despite such evidence, the concept of “cognates” usually holds only between languages from the same family (like Romance languages), and can hardly, if ever, helps the communication process between speakers of languages from different language families.

It is not known exactly when natural languages were first used by man. Speculations [16] have linked the first usage to be between two million years ago, during the time of Homo habilis, to four thousand years back, during the time of Cro-Magnon. It is widely understood that the earliest existence of languages came in the form of speech, before any organization into symbols, and later, grammars. That said, some early languages were formed on a set of symbols available, such as the Ancient Greek language of the Archaic and Classic periods. Due to the ephemeral nature of speech, we do not have clear ideas of language usage in the early days, otherwise, such findings can facilitate our understanding of evolution of languages in association with the perceptual and cognitive system of human brains.

However, as time progresses, many new languages have surfaced. According to recent studies [11], [12], there are more than 7,000 languages spoken worldwide. Of these, there are many which do not develop into formal language systems e.g. many dialects spoken by minority tribes in China. The reasons why we are experiencing multitudes of such languages today could perhaps find its roots in geographical, cultural, racial or even religious differences - people develop different spoken forms in their own exclusive groups to encode information and ideologies for internal understanding.

For better or for worse, this diversity impacts our lives on a daily basis. While it is beyond doubt that having different languages increases our appreciation for each other's differences, and provides a framework for discovery of the human mind's conceptual system of relating entities using their world knowledge, people nowadays are beginning to feel the strain of not being able to communicate smoothly in many situations because of "language barriers".

Universal communication still represents one of the long-standing goals of humanity - borderless communication between people, regardless of the language they speak. In his bestseller *The World is Flat*, Thomas Friedman points out that the "breaking down of political, cultural and trade barriers in globalization, as well as the exponential technical advances of the digital revolution, have made it virtually easy for people to engage in any transactions with billions of others across the planet". Such is the effect of globalization that no man can live without significant interaction with the outside world. Clearly, the result is a greater need for a more seamless means of communication among people who share different languages.

No doubt, sharing a common language between the speaker and the listener provides a direct way of expression, and this is still by far the best way to communicate. However, this method is not without its limitations. Particularly, a person's ability to grasp many languages is limited. Even learning to use a language effectively and efficiently requires many years of training.

In this thesis, we seek to augment our present verbal/written communication using a new paradigm : pictorial representations that have been proven to possess the ability to convey meanings encoded in short sentences across “language barriers”. For instance, *The house has four bedrooms and one kitchen* in its pictorial form (Figure 1.1) generates *universal* understanding among people who speak different languages.

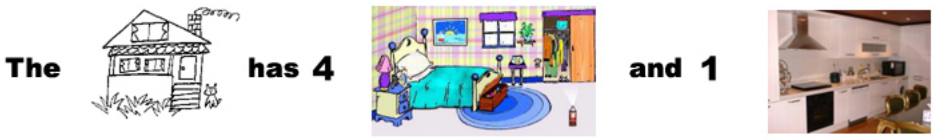


Figure 1.1. A pictorial translation for “*the house has four bedrooms and one kitchen*”.

Surprisingly, studies [15] have shown that only up to an estimated 10 percent of our communication is verbal. The “unspoken language” is our body language. This refers to our behaviors while making the speech, such as smile, gaze, attention span, attitude, arms movement, head shaking etc, all perceived by our listener through their visual system. Arguably, by not using our eyes during communication, we potentially lose 90 percent of the content of information !

In fact, before any verbal communication is established, man had used proto-linguistic or non-linguistic means to make himself understood. This evidence strongly suggests that visual representations of information is very helpful to a non-negligible extent, and they require minimal learning in most instances. Most importantly, our aim is to establish, through practical situations, that they are universal.

In addition to enabling communication across “language barriers”, the ability to encode information using pictorial representations has other potential benefits, such as language learning for children or for those who study a second language, or understanding for people with a language disorder. So far, natural language processing (NLP) research is largely restricted to syntactic and semantic intra-language domains. If established, such a visual information system can also be used to facilitate cross-language research, and open up a new branch of potential benefits for NLP researchers, artificial intelligence (AI) researchers alike. It can also bridge the gap between work in image processing and NLP, thereby facilitating the application of methods learned in one domain to the other.

The thesis is organized as follows :

In chapter 2, i shall present some related work done in the area of words and pictures research. In chapter 3, i introduce the various resources and tools used in our project. Specifically, in chapter 4, i elaborate on PicNet, a web-based system for augmenting semantic resources with illustrative images using volunteer contributions over the Web, which is the main tool for our research. In chapter 5, i provide our hypothesis, conduct experiments and discuss the results. Also, i supply a few examples to illustrate our point. In chapter 6, i explain the automatic pictorial translation system. Finally, in chapter 7, i conclude our findings, summarize our contributions and discuss rooms for future work.

CHAPTER 2

RELATED WORK

Research that simultaneously addresses words and pictures has been carried out in different fields of science, including cognitive science, image processing, computational linguistics and visual languages. For the purpose of situating our work in context, we shall highlight some of the works that have contributed significantly to this line of research.

2.1. Cognitive Science

The meaning of a sentence is encoded in each of its component words. In order to understand a sentence, all meanings of the individual words must be retrieved and combined. Do humans achieve this retrieval of meaning through a lexicon that is part of a linguistic system, or is the meaning stored as part of a general conceptual system in the brain ?

Early research efforts in cognitive science and psychology [25] have proposed two such similar hypotheses to empirically determine how word meanings are processed. The authors used *rebus sentences* in which a concrete noun is replaced by a pictured object. Such sentences consist of 10 to 15 words each, picture(s) inclusive. They are shown using rapid serial visual presentation (RSVP) to forty subjects. The rationale of using RSVP at a rate of 10 or 12 words per second is to present sentences so quickly that a delay in encoding the picture into a required form (e.g. silently naming it to help establish semantic link between the word before and after) would be highly disruptive to the subjects, and hence producing results that bias towards either of the hypotheses.

In all experiments performed, there was no significant delay in understanding of *rebus sentences* compared to all-words sentences. Accuracy wise, there was no consistent deficit in their interpretations. In fact, the speed of understanding and accuracy of comprehension or immediate recall remains the same regardless of the position of the picture (front, middle or end of the sentence), nor did it matter whether there was one picture or two pictured object replacing concrete nouns.

Clearly, the experimental results do not support theories that suggest word meanings are placed in a specialized lexical entry. On the contrary, the lexical representation of a noun merely points to a general, nonlinguistic conceptual system in human brains where the meaning of a sentence is constructed. This finding points to a *truth* that our comprehension of real-world entities is not restricted by information encoded in any language system. It also opens up a possibility that humans can communicate with one another through non-linguistic means.

2.2. Image Processing

Currently, the best search engines on the Internet (Google, Yahoo, Alta Vista) thrive on a high precision and recall of information requested by Web users. However, this success has been largely limited to Information Retrieval (IR) in the natural language domain, while searching for pictures still gives low precision and poor recall. The reason behind this limitation is that most pictures are processed using the captions that label them. Since these captions are almost always packed with a high degree of noise, better searches can only result from searching beyond the captions. The content of each picture must be analyzed meaningfully, then tagged with correct words that describe it, before it can be retrieved using traditional textual IR methods.

In a line of work relating words to pictures [4], the authors presented a new approach for modeling multi-modal sets, using segmented images with associated text. By learning from the joint distribution of image regions and words, many applications can be yielded. These include predicting words associated with whole image (auto-annotation) and corresponding to particular image regions (region naming). Due to the difficult nature of applying data mining methods to collections of images, learning the relationships between image regions and its semantic correlates (words) proves to be an alternative method of multi-modal data mining.

In another effort to improve object recognition [13], researchers successfully showed that by automatic captioning using a novel graph-based approach (GCap), words may be reliably predicted from images. The assessment was done on the "standard" Corel image database where GCap outperforms recent, successful automatic captioning methods by up to 10 percentage points in captioning accuracy. This method is fast and scales well with its training and testing with time linear to the data set size. Besides, it requires no user-defined parameters, nor other tuning, which is in contrast to linear/polynomial/ kernel SVMs, k-means clustering etc.

These advances in words and images research can lead to better image recognition, and hence produces higher image retrieval accuracy. In our proposed paradigm of using pictures to replace words, it is imperative that good quality images are retrieved from search engines based on all-words search. We will discuss more on the search engines used in our project in chapter 4.

2.3. Computational Linguistics

Traditionally, word sense disambiguation (WSD) has been a well-studied problem in computational linguistics. Given a word in a sentence, the task is to determine which sense of the word (with multiple senses) is used in the context of that sentence.

Consider the word *bank*. Examples of different senses include *piggy bank* (a container for keeping money), *river bank* (a slope of land besides a body of water), *Wells Fargo Bank* (a financial institution), *snow bank* (a long ridge or pile) etc. Picking the correct sense can be potentially challenging because of metaphorical or metonymic meanings that makes discrimination of closely related senses difficult. Also, there is the issue of interjudge variance. WSD systems are usually compared to a benchmark sense-tagged corpora by humans. However, even when creating this benchmark, decisions to arrive on which sense to use for a given word varies across human judges.

With the same spelling for each word to be disambiguated, it is hard to adopt a purely natural language based approach and expect a very good result over the "most common sense" method (which selects the most frequently used sense), usually used as a base line. In an innovative effort [5], experiments revealed that using pictures can actually help disambiguate words, while the reverse is also true. Starting from a learned set of pictures associated with words, co-constructed meanings can be established from these two different representations of the same entity. The images are then combined with sophisticated text based word sense disambiguation methods to perform disambiguation tasks over a subset of Corel image database with three to five keywords per image. The results show that this technique is superior to using pictures or text based methods alone for disambiguation. The hypothesis governing this observation is that properties implicit in one representation may be more explicit and therefore more extractable. Given a large copora for training, the relationships between these two can be learned, and hence pictures can be used to provide a non-negligible improvement over WSD tasks.

In our project, we require a specific sense of the word to be identified prior replacing it with a picture. This step is necessary to produce accurate pictorial translation for a sentence. We rely on the help of a state-of-the-art word sense disambiguation tool to achieve this process. More details would be given in chapter 3.

2.4. Visual Languages

Visual Language is an expression system involving the use of visual objects to express our thinking and feeling. It stems from the pioneering work of Rudolf Arnheim to studies by Robert Horn and includes the use of a visualizing method called *active imagination* developed by Carl Jung [26]. Built on the proposition that we can "draw" our thinking as well as verbalize it, a Visual Language may contain words, images, and shapes.

Particularly, there is a branch of visual languages called *iconic languages*, whose visual sentences consist of a spatial permutation of icons. Each icon bears a unique (or sometimes multiple) meaning in the vocabulary set of icons used in Iconic Language. In human-computer interaction, the iconic language normally has a limited vocabulary set with specific application domain such as database access, form manipulation and image processing. To facilitate the design of such iconic languages, a design methodology was devised [8] based on upon the theory of icon algebra, allowing for a flexible derivation of the meaning of iconic sentences.

In the human-human interaction, there are also iconic languages used, especially in augmentative communication by people with speech disabilities. Much work also has been done in the area of augmentative and alternative communication regarding the use of visual-graphic symbol acquisition by pre-school age children with developmental and language delays. In their findings, the authors [1] concluded that the acquisition of a language requires an individual to organize the world into a system of symbols and referents. However, learning the relationship between a symbol and referent can be difficult for a child with serious intellectual disability and language delays. The complexity and iconicity of a symbol becomes an important issue in the decision of what medium to use for teaching languages. By using an observational experiential language intervention, they are able to study the effects of four pre-schoolers with developmental and language delays to acquire the meanings of Blissymbols.[1] and lexigrams. The results confirmed the findings that even children with such disabilities are able to acquire language skills through visual representations, although performance varies according to the participants' comprehension skills.

Commercial products with visual interfaces[2] have also been marketed with success. They are used for augmentative communication for people with physical limitations or speech impediments, with iconic keyboards that can be touched to produce a voice output for communication augmentation. Also related to some extent is the work done in visual programming languages, where visual representations such as graphics and icons are added to programming languages to support visual interactions and to allow for programming with

[1] Blisssymbols is a symbolic, graphical language that is currently composed of over 3,000 symbols

[2] http://www.amdi.net/

visual expressions. Additionally, there are also many pictorial dictionaries[3] that boost the quick acquisition of a new language skill, through the use of word/image associations. Specific pictorial references based on architectural[4] and medical domains[5] are also available; these are excellent learning aids for the various professionals in their fields.

Research done in the different fields of Science has lend strong credibility to our hypothesis that pictures can replace words not only in their individual meanings, but entire sentences can potentially be translated into pictures, yet generating the same level of understanding desired. The mutual relationship between words and pictures means that under certain context, they are interchageable while maintaining the semantic structure in the sentence.

[3] http://www.pdictionary.com/, http://web.mit.edu/21f.500/www/vocab-photo/

[4] http://architecture.about.com/

[5] http://www.nlm.nih.gov/medlineplus/encyclopedia.html

CHAPTER 3

BACKGROUND ON RESOURCES

The pictorial translation paradigm starts with pre-processing an input sentence and ends with a pictorial representation of the sentence given as the output. Extensive resources and tools are needed to complete this process. Prior to any processing on the word, knowledge must be gathered about its meaning, morphology and its relation to other words; the use of Wordnet as a machine readable dictionary serves this purpose. Next, a part-of-speech tagger is needed to tag each word. This is done using Brill's tagger, a state-of-the-art rule-based tagging system. When presented with multiple senses of the word, it does not suffice to know only the part-of-speech - the exact sense of the word must be selected. We achieve this using another high-performance word sense disambiguation tool called SenseLearner. Finally, once a word has been labeled with its part-of-speech and meaning, we need to replace it with a suitable picture. PicNet is an important resource which we would discuss in full details in chapter 4. In the sections below, we provide details on Wordnet, Brill's tagger, SenseLearner, and Machine Translators evaluations.

3.1. Wordnet

Wordnet®[1] [21] is an online semantic lexicon for English language, its creation being inspired by current psycholinguistic theories of human lexical memory.

Wordnet is distinctively different from a dictionary or a thesaurus. In a dictionary, words are ordered according to an alphabetical order, while their meanings are scattered randomly throughout. In a thesaurus, words are grouped together semantically at the expense of their alphabetical order. Wordnet attempts to combine the best of both worlds with the construction of a highly searchable lexical list with each entry belonging to a *synset*, which

[1] Wordnet online lexical reference system, Princeton University, NJ, http://wordnet.princeton.edu/

is a set of synonymous words or collocations (a collocation is a sequence of words often used together to show a specific meaning e.g. “car pool”). The meaning of a synset is further clarified with a short defining *gloss*. Each synset with its set of words and gloss represents a single conceptual entity and forms the most basic constructing unit for Wordnet.

In reality, Wordnet is not just seen as a vast collection of synsets. Semantically, there exists meaningful links between synsets. One example is the important *antonym* relationship which denotes a synset as having opposite meaning to another.

These relationships are modeled in a way that reflects the organization of a lexicon in the human memory. Together, their existence form a web of semantics where there is a pointer from each synset (a meaning, or a single conceptual entity) to another governing the type of relationship held. This richness of information and its semantic links imply the suitability of Wordnet for use in Natural Language Processing and Artificial Intelligence applications.

WordNet also provides the *polysemy* count of a word, which is the number of synsets that contain the word. When a word appears in more than one synset (i.e. more than one sense, or meaning, or conceptual entity), it implies that some senses are much more common than others. Wordnet uses *frequency scores* to quantify this phenomenon. In a sample corpus, all words are semantically tagged with the corresponding synset, after which a count was given on how often a word appeared in a specific sense.

As of 2005, Wordnet database contains more than 150,000 words organized over 115,000 synsets for a total of 203,000 word-sense pairs. In compressed form, it is approximately 12 megabytes in size. Table 3.1 shows the number of nouns, verbs, adjectives, and adverbs defined in Wordnet 2.0, and the number of synsets for each of these parts of speech.

3.1.1. *Semantic Relationships between Noun Synsets*

Of two important relationships among noun synsets are the Inheritance and Part-Whole relationships. The Inheritance relationship simply means features are inherited from one word to the other. Figuratively speaking, words with inheritance links are ordered on a

Part of Speech	Words	Synsets
Noun	114,648	79,689
Verb	11,306	13,508
Adjective	21,436	18,563
Adverb	4,660	3,664
TOTAL	152,050	115,424

TABLE 3.1. Words and synsets in Wordet 2.0.

hierarchical basis, with the lower levels inheriting from the higher levels. Wordnet classify this type of relationship into the *hypernym* relationship, which states that X is a kind of Y if Y is a hypernym of X, and conversely, the *hyponym* relationship, which states that Y is a kind of X if Y is a hyponym of X. When two words share a common hypernym, we call them *coordinate terms*. Hence Inheritance can also be thought of as "IS A" relationship. For instance, a "dog" is a "canine", a "canine" is in turn a "carnivore", a "carnivore" is in turn a "placental" and so on. This is shown in Figure 3.1.

Note that the hypernym-hyponym relationship is transitive, meaning that a "dog" inherits from "mammal" i.e. a "dog" is also a "mammal" if Y is a hyponym of X. Also, the relationship is one-to-many, meaning that a "dog" can only be a "mammal", not a "reptile" at the same time, but besides "dog", a "cat", a "pig", a "duck" can all be "mammals". They are coordinate terms.

Part-whole relationship indicates a "PART OF" relationship. Intuitively, a "hand" is a part of a "body" , and hence qualify for the part-whole relationship. We call the "hand" a *meronym* of the "body", and conversely, the "body" is a *holonym* of the "hand".

Part-whole relationships are similar to Inheritance relationships in their hierarchical structure and transitivity. The two type of relationships can be combined to form a composite

```
dog
 ⇒ canine, canid
    ⇒ carnivore
       ⇒ placental, placental mammal, eutherian, eutherian mammal
          ⇒ mammal, mammalian
             ⇒ vertebrate, craniate
                ⇒ chordate
                   ⇒ animal, animal being, beast, brute, creature, fauna
                      ⇒ organism, being
                         ⇒ living thing, animate thing
                            ⇒ object, physical object
                               ⇒ physical entity
                                  ⇒ entity
```

FIGURE 3.1. A Wordnet "is a" relationship for *dog.*

relationship, as in if X is a hyponym of Y, and W is meronym X, and Z is a meronym of Y, then W can be a hyponym of Z.

3.1.2. *Semantic Relationships between Verb Synsets*

Verb synsets also exhibit hypernym-hyponym relationships between them. A clear instance of such a relationship is the verb "walk". "stagger", "trudge", "stride" are all hyponyms of the hypernym "walk".

A relationship exclusive to verb synsets would be *troponym.* To understand troponymy, we first visit *entailment*, a concept well-defined for propositional logic. When X entails Y, we state that under no conceivable state of affairs, there exists a situation of X is true Y is false, and vice-versa. Now, if we say "snore" entails "sleep", there is no way whatsoever to state confidently that "sleep" entails "snore" too, and hence the relationship is unilateral. Troponymy, thus, specify every verb X entails a more general verb Y, if X is a troponym of Y.

The causative relation relates the “cause” (in the word “display”) to “effect” (in the word “see”). This type of relation is transitive; if X causes Y, Y causes Z, then we conclude that X causes Z.

Besides having semantic relations in a category, there are also semantic relations connecting different categories. For instance, an adjective modifies an attribute, resulting in a link between the adjective to the synset containing the attribute. An adverb may link to an adjective from which it is derived.

3.2. Brill’s Tagger

Manual annotation of part-of-speech on large corpora is a painful process, hence the need to automate the process. Early research efforts has produced mostly simple Markov-model based stochastic taggers. These type of taggers assign a sentence the tag sequence that maximizes *Prob(word | tag)* * *Prob(tag | previous n tags)*, with the required probabilities estimated from a manually tagged corpus. Although stochastic taggers produce high accuracy for tagging, they have the disadvantage of capturing linguistic information indirectly in the form of large statistic tables.

In this section, we introduce a different kind of part-of-speech tagger that uses a rule-based approach, but yet produce tagging accuracy that is comparable to that of stochastic taggers. The advantage of using a rule-based tagger is that relevant linguistic information is captured in a small set of simple non-stochastic rules.

Transformation-Based Error-Driven Learning is a paradigm that has been successful in solving a variety of natural language processing problems ranging from part-of-speech tagging to syntactic parsing. The learning process is depicted in Figure 3.2. An unannotated text is passed through the initial-state annotator, which can have a range of complexity from labeling random structure to assigning the output of a sophisticated manually created annotator. After the initial-state annotation, the text is compared with the *truth*, which is specified in a manually annotated corpus, and transformations are then learned that can be applied back to the annotated text to make it look more like the *truth*.

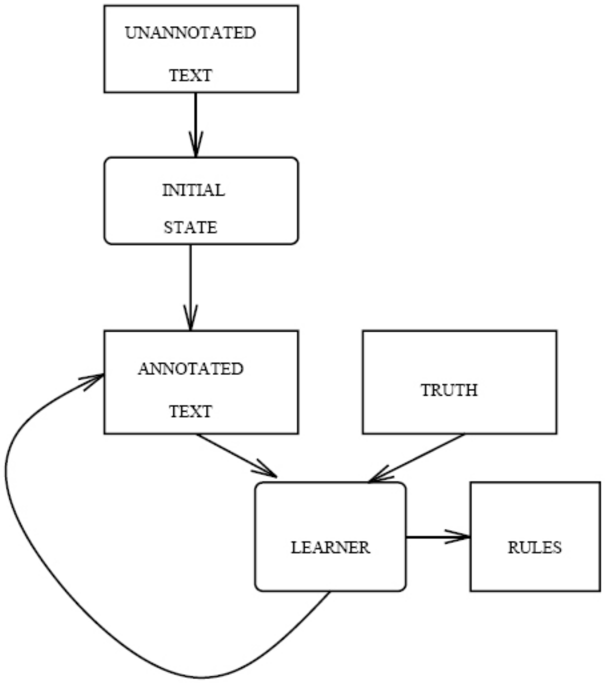


Figure 3.2. Transformation-based error-driven learning.

In Brill's tagger, the idea is to search for the transformation in each round which has the highest score; this transformation is added to the list of ordered transformations, and the training corpus is updated by applying the learned transformation. Specifying an instance of transformation-based learning involves stating the following (1) the initial state annotator (2) the space of transformations allowable for the learner (3) the scoring function used to compare the corpus to the truth and choosing a transformation. Rules are obtained in the list of ordered transformations and now can be applied to the output of initial state annotator, one by one.

Unlike stochastic taggers, Brill's tagger captures important relationships between words directly into a set of explicit rules. Contextual transformations referencing relationships between a word and the previous word, or between a word and the following tag, and such others, are addressed. Below we list two examples of learned lexicalized transformations.

(1) From *preposition* to *adverb* if the word two positions to the right is *as*.
(2) from *non-3rd person singular present verb* to *base form verb* if one of the previous two words is *n't*.

To illustrate the usefulness of (1), we consider the phrase *as tall as*. Using stochastic taggers, the following annotation would be produced :

as/PP *tall*/JJ *as*/PP

According to Penn Treebank tagging style manual, the first *as* is tagged as an adverb while the second *as* would be tagged as a preposition. In the initial annotation for Brill's tagger, the first *as* is also mistagged as a preposition, because its most frequent tag encountered in the training corpus is also a preposition. However, the first transformation later would correct this mistagging.

Another strength of Brill's tagger lies in its ability to tag unknown words with a high accuracy. Usually, part-of-speech taggers resort to a back-off method in using the most frequent tag seen in the training corpus, when deciding on the tag for an unknown word in the test corpus. The logic is that once a transformation-based tagger can assign the most likely tag for an unknown word with high accuracy, then contextual rules surrounding this word can be used to further improve the overall frequency. The process for prediction starts with the initial state annotator which naively labels the most likely tag for unknown words as proper noun if capitalized and common noun otherwise[2]. Then, a set of allowable transformations is applied, which exploit the morphology of words (e.g. common suffixes, prefixes)

[2] A rule must be learned of the form : change tag to proper noun if the prefix is "E", since the learner is not provided with the concept of upper case in its set of transformation templates

to derive the most likely tag.

Brill's tagger obtains competitive results with stochastic taggers in both known and unknown words tagging. For known words, When trained on a mere 600K size corpus, with just 267 contextual rules, Brill's tagger scores an overall accuracy of 97.2%. This is in contrast to stochastic traninig which requires 1 million size training corpus with 10,000 contextual probabilities to produce an overall accuracy of 96.6%. For unknown words, only 148 rules were learned, yet producing a very competitive 85% accuracy.

3.3. SenseLearner

In word sense disambiguation, the task is to assign a suitable meaning to a polysemous word in a given context. This process forms an important part of any application requiring knowledge about meanings, such as machine translation, knowledge acquisition, question answering, information retrieval and the likes. Most current word sense disambiguation systems rely on supervised learning where each word is tagged with a part-of-speech and later transformed into a feature vector for automatic learning. One drawback of such a supervised learning algorithm is that a large corpus of sense-tagged data containing those candidate words must be available, and the accuracy is very much dependent on the quantity of these data available.

In this section, we discuss SenseLearner [18], a minimally supervised algorithm that attempts to disambiguate all content words in a unrestricted text. SenseLearner takes as input a small set of sense-tagged data for learning (and hence is considered minimally supervised), and creates generalized semantic models which can be applied to disambiguate any word without the need to use a separate classifier.

3.3.1. *Algorithm*

SenseLearner fits well as a WSD tool for our project because it is minimally supervised (requires little training data sets), general (able to disambiguate most words in our texts) and efficient (perform disambiguation in real time). The data sets used for training are drawn from SemCor [22], which is a corpus of manually sense-tagged words by experienced lexicographers.

In the algorithm, the precondition is a raw text with words to be disambiguated. The postcondition is the same text with word meaning annotations for all the open-class words.

Initially, preprocessing is done to the raw text, this involves tokenization where meaningful words are identified. Next, these tokens are assigned with part-of-speech tags; collocations (which are sequences of words that appear together often to denote a compound meaning e.g. *car pool*) are picked out using a sliding window approach. Also, in this phase, name entities are identified. Note that SenseLearner only identify persons, locations, and groups which are the only specific named entities in SemCor.

Following preprocessing is a phase of building semantic models for all predefined word categories. All words in a category are either syntactically or semantically similar to one another. Additionally, word categories can be refined into further granularities. Starting from the most general, we may have a semantic model that handles all *nouns* in the test corpus. Using a similar mechanism, we may build another semantic model, this time being more specific, that handles all *verbs* with at least one of senses of type *move*. On further finetuning, we would have a most specific model that handles one word at a time. After being defined and trained, these semantic models are applied to the test corpus for words disambiguation. Various models addressing part-of-speech and collocations that are currently implemented in SenseLearner are detailed as follows :

Noun Models

modelNN1: A contextual model that relies on the first noun, verb, or adjective before the target noun, and their corresponding part-of-speech tags.

modelNNColl: A collocation model that implements collocation-like features based on the first word to the left and the first word to the right of the target noun.

Verb Models

modelVB1 : A contextual model that relies on the first word before and the first word after the target verb, and their part-of-speech tags.

modelVBColl : A collocation model that implements collocation-like features based on the first word to the left and the first word to the right of the target verb.

Adjective Models

modelJJ1 : A contextual model that relies on the first noun after the target adjective.

modelJJ2 : A contextual model that relies on the first word before and the first word after the target adjective, and their part-of-speech tags.

modelJJColl : A collocation model that implements collocation-like features using the first word to the left and the first word to the right of the target adjective.

Defining New Models

Arising needs to define new semantic models is addressed in the current version of SenseLearner. A template that cover subroutines to create such a model is provided. After constructing the model, it can be trained in the same way like other predefined models.

Whenever a target word appears in the test corpus but is not contained in the training corpus, or it is not covered by any of the semantic models, a back-off method, which employs the most most frequent sense in Wordnet, is used.

During training, a feature vector, with features specific to each model, is constructed for each sense-tagged word. Feature vectors are added continuously as training progresses, and grouped under each model for each word. The label of each vector is in the form *word#sense*. Following this stage, learning starts for all test examples, using Timbl memory based learning algorithm [10]. Note that again, each learning is started per semantic model per word, and word sense predictions are made here. A predicted word and its sense is annotated only when all models agree on the same prediction, otherwise there is no annotation at this point - it shall be annotated later.

The SenseLearner algorithm is shown in Figure 3.3

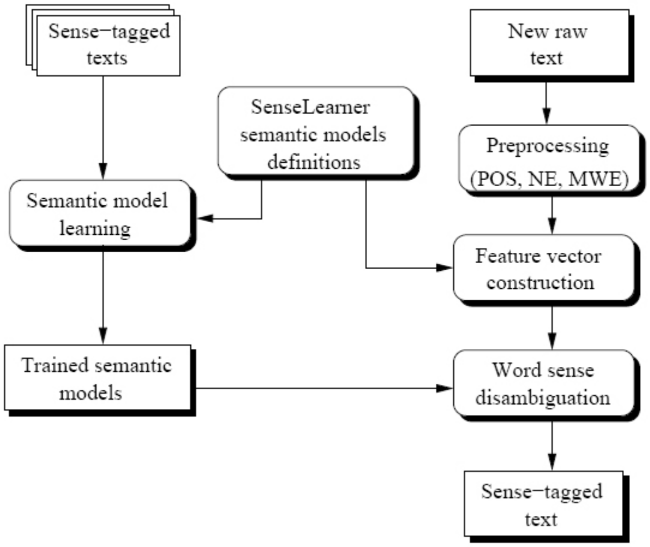

Sense−tagged texts
New raw text
SenseLearner semantic models definitions
Preprocessing (POS, NE, MWE)
Semantic model learning
Feature vector construction
Trained semantic models
Word sense disambiguation
Sense−tagged text

Figure 3.3. Semantic model learning in SenseLearner.

SenseLearner is evaluated is used to disambiguate Senseval-2 and Senseval-3 English all words data sets, each data set consisting of three texts from the Penn Treebank corpus annotated with Wordnet senses. The overall accuracy is 66.2% for Senseval-2 and 64.1% for Senseval-3, which is found to be a significant improvement over the simple yet competitive baseline that chooses by default the most frequent sense. SenseLearner also compares favorably with the best published results over the same data sets.

3.4. Machine Translation

Machine translation achieves the task set up in natural language translation through the use of computer software and/or hardware. By definition, natural language translation is the process of establishing an equivalent of a text or speech in another language. The process includes : (1) Decoding the meaning of the source text; and (2) Encoding the meaning in the target language.

Seemingly simple, these two steps entail complex cognitive functions. To decode the meaning of the source text in its entirety, the translator first interprets the source text in depth using all its features - grammars, syntax, semantics and sometimes the culture of the native speakers of the language. To complete the translation, a similar level of understanding of the target language and its components must be present. Arguably, since the translator needs to actively express in lexical terms, the knowledge of the source target must be deeper than that of the source language. Hence in many cases, translators often translate into a language of which they are native speakers.

3.4.1. *Approaches*

Machine translation is based on a set of linguistic rules rendered into machine readable forms as a means of guiding the computer in producing the most suitable translation. Crudely speaking, it performs simple atomic substitution of one word for the other in the target text. Using corpus techniques, better understanding can be achieved of the differences in linguistic typology, phrase recognition and translation of idioms to improve accuracy in the translation.

In a nutshell, several methods have been developed and used for machine translation, of which rule-based and statistical methods are the most dominant.

3.4.1.1. *Rule-based Methods.* Generally, rule-based methods [2] create an intermediary representations by parsing the source text, and uses these to generate the target text. By nature of its intermediary representation, this approach is usually described as interlingual machine translation, or transfer-based translation. The success of rule-based methods depends largely on the existence of lexicons with extensive morphological, syntactic, and semantic information, and large sets of rules.

3.4.1.2. *Statistical Methods.* Statistical-based methods [23], also referred to as example-based methods, require the use of parallel corpora, such as the Canadian Hansard corpus and the English-French record of the Canadian parliament to generate translations. The advantage of using these methods is that no manual lexicon building or laborious rule-writing is needed. In fact, very competitive results can be achieved where parallel corpora exists between the source and the target texts.

3.5. Machine Translation Evaluation

Manual and automatic evaluation schemes are proposed over the years, with human judgment to be the oldest and most reliable. Automatic evaluations are suitable for facilitating tasks of evaluating large translation sets efficiently. However, they do not prove to be reliable judges due to paraphrasing and synonyms anomalies in the translated text. Usually, a mixture of both manual and automatic evaluations is employed to give an overall standard of quality of the machine translation.

3.5.0.3. *NIST.* BLEU (acronym for Bilingual evaluation understudy) [24]. The NIST score is based on the BLEU metric. It is an automatic method for evaluating machine translation. The quality of translation is indicated as a number between 0 and 1 and is measured as statistical closeness to a given set of good quality human reference translations. Therefore, it does not take into account translation intelligibility or grammatical correctness.The metric works by measuring the n-gram co-occurrence between a given translation and the set of reference translations and then taking the weighted geometric mean. BLEU is specifically designed to approximate human judgment on a corpus level and can perform badly if used to evaluate the quality of isolated sentences.

3.5.0.4. *GTM.* GTM (acronym for General Text Matching) automatically measures the similarity between texts and can be used to evaluate machine translations. [17] Standard evaluation measures like precision, recall and F-measure are incorporated into GTM to give a combined rating between 0 and 1, which is the degree of *similarity* between the translated text and a specific reference text. This protocol of combining metrics is based on the experimental findings different measures relates differently to different corpora and requires specification of tuning parameters in each metric.

3.5.0.5. *Humans.* Traditionally, humans assess quality of translated text with references to both its *adequacy* and *fluency*. Adequacy is judged by comparing the content of each translated text with a high quality reference translation. Typically, spelling errors, grammatical mistakes and sentence structure do not contribute to scoring in this adequacy aspect. Meaning of the translated text in its entirety is considered. In the fluency part, the human checks for correctness in grammar structures and elegance of phrasing in the translated text. Usually, a scale of 0-5 is given, 0 being the worst assessment and 5 being the maximum score.

CHAPTER 4

PICNET

This chapter describes in detail PicNet [6], a Web-based system for augmenting semantic resources with pictorial representations by using contributions from Web users. PicNet serves to build a rich knowledge-based by combining word/picture associations to the effect of capturing world concepts which are used in our system for generating pictorial translations for simple sentences. We discuss the motivations behind the PicNet project, as well as some of the issues that pertains to the construction of such a knowledge-base.

4.1. Motivation for PicNet

The motivation behind PicNet is the desire to build a semantic network that extends Wordnet to capture both visual and linguistic representations of a concept. Wordnet by itself is a lexical knowledge-base confined to English speakers, hence extracting the rich semantic information contained in Wordnet to benefit non-English speakers can be achieved through the use of a visual/linguistic association. Such an association is treated as the smallest cognitive unit of information in PicNet.

Children who are preliterate can also make use of the word/image associations in PicNet to the effect of picking up the connection between two such representations faster, and hence master the linguistic representations in a shorter time. People who are learning to pick up English as a second language may find PicNet a favorable learning aid. In fact, another goal of PicNet is to become an international language-independent knowledge-base that facilitates exchange of information across language barriers between people who do not share common languages. This leads to a potential use of PicNet as a multilingual dictionary, where a single pictorial representation may be linked to several linguistic representations.

In another inspiration, PicNet can be seen as a tool to bridge research between language processing and image processing. As an example, the explicit word/image associations in PicNet and hierarchical semantic links between such associations may be deployed to improve quality of image retrieval systems and/or classification. Conversely, image content analysis can also be used to help language processing tasks, as explained in chapter 2. Using word/image associations may also lend benefits to information extraction, information retrieval, named entity recognition and so on.

4.2. Encoding Word/Image Associations

Initially, PicNet focuses on the use of concrete nouns. These are nouns that relate specifically to a concrete entity in the real world (e.g. *water lily*, *bulldog*). There exist other types of entities (e.g. general, abstract nouns) and interactions between such entities that conforms a different form of visual representation. Taking an example like *flower*, it suggests only a general lexicalization and does not imply a particular type of *flower*, like *sunflower*. Representing *sunflower* entails *flower*, this understanding is implicit. However, explicit representation is desired, and hence the latter requires a more generic form of visual representation. For this and other abstract nouns, PicNet suggests the use of various mechanisms to visualize them. In the case of general nouns, an *instance morphing* method can be used to create the illustration of a general concept - *car* may be implied by sequencing Honda, Nissan, Ford, Mercedez in a cycle. To relate the information contained in a abstract word, such as *philosophy*, employing the use of famous philosophers i.e. Socrates, Pilate, Confucius, again, cycling in a sequence can be a good method. Attributes of nouns can be inferred from a collage of pictures of those nouns possessing such attributes. Lastly, verbs in action can be represented as an entity carrying out that action. The meaning of *drink* can be visualized as a picture of a man drinking a glass of water. The above-mentioned mechanisms require more elaborate testing to be regarded as useful. Note, also, that PicNet does not restrict the number of images that can be associated with a word. It is not likely that one image on its own is able to fully conceptualize a given word, rather, collections of such images from different users can provide a larger scope of overall understanding to any single user.

4.3. Resources

The main objective of PicNet is to build a large knowledge-base by leveraging the help of Web-users contributions. The intuition is that all language speakers are experts when it comes to relating a linguistic representation to its pictorial correlate. In order to build a database of word/image associations, a lexical database and a pictorial database are required foremost. With these in place, it is relatively easier for Web-users to contribute to PicNet, by relieving them of the burden to look for such resources themselves. Their main task then is to establish the links between the semantic nodes of each database. To do this, there are a variety of ways to engage users in an interactive way, which would be elaborated in the next section. Two important resources used in PicNet are described below.

4.3.1. *Wordnet*

The use of Wordnet 2.0 (explained in detail in chapter 3) constitutes an important semantic network of information from which users can draw from. Concepts, each of which bears a definition and identified by a synonymous group of words, can be interlinked to one another. These semantic links such as hypernymy-hyponymy (IS-A), meronymy/holonymy(HAS-A) can be exploited in meaningful ways the network of pictures built in PicNet.

4.3.2. *Image Search Engines*

Automatic collection of images is performed through PicSearch[1] and Altavista[2] websites. At present, more than 73,000 images have been collected. Manual validation on these images is done by volunteers on the Web. Automatic validation may be possible, however, not implemented in the current version of PicNet. This automatic collection of picture is

[1] http://www.picsearch.com

[2] http://www.altavista.com/image/

rated at 61% sucessful[3]. Though this system proves very efficient at collecting images related to concrete nouns with precise definitions, there are other concerns, such as its inability to differentiate between senses of the same word, and searching based soley on the textual description rather than relying on image content analysis for a more accurate hit.

4.4. Activities in PicNet

In any Web-based data collection system, a major goal to maximize users participation in order to capture the limitless amount of knowledge that can be tapped from those users. PicNet seeks to provide engaging activities and competitive games that maintain users in an motivated state while collecting valuable data. Maintaining the integrity of the system is crucial to the success of the system. This is achieved through the use of an administrator facility which allows a superuser to disallow malicious uploads from any Web users. The superuser is also allowed to undo any earlier transactions where necessary.

4.4.1. *Searching*

A user can search PicNet for word/image association based on a word she enters. In this case, all synsets are searched and a list of synsets containing the word is returned, together with a picture for each synset. This is because a word may coexist in several synsets, with each defining a different concept. If more than one picture exists for a synset, only the top ranked picture is returned. At this point, the user may be dissatisfied with the picture and elect to upload a new picture for that particular concept. Note however, that this newly uploaded picture is checked against integrity constraints before it is integrated into PicNet system. The user may also comment on the quality of the word/image association by giving a rating on it - another activity that would be covered later. The system is capable of performing wildcard searches, meaning that if the user enters *tiger*, other words such as *tiger lily* and *tiger shark* may be returned.

[3] 44% of the images were good matches, 17% were a near match with some common properties exhibited by the corresponding synset

4.4.2. *Donating Images*

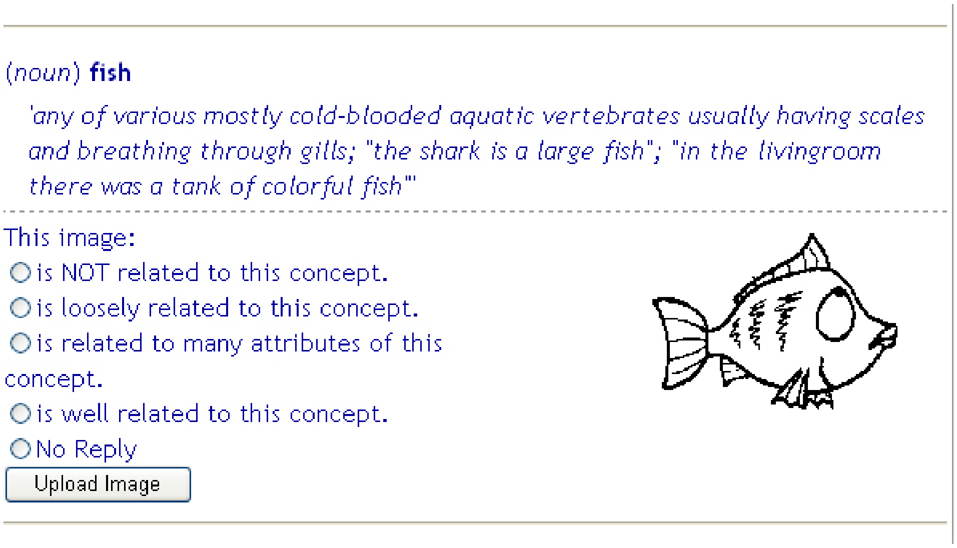
A user can elect to upload images for use in PicNet. Again, such images are not immediately integrated into the system - it will be validated first.

4.4.3. *Free Association*

A user is shown a picture that is randomly selected from PicNet's database. A synset may or may not have been tagged to this picture, this is unaware to the user. The user is asked to enter a word that describes the picture. Based on the word entered, more refined definitions are produced following a search in Wordnet. The user then chooses from one of these to be used for a new synset/image association (given an automatic + 3 point vote). It is probable that different users label different words for the same given picture. This is expected, and serves to validate the relationship between synsets (such as hypernym-hyponym). It is even possible that such a relationship is novel and not yet before recorded in Wordnet. Being presented a more specific instance of a word and its picture, a user can provide a naming that is in fact more general. When presented with a picture of *mule deer*, the user may thought of it as merely *deer*. Seemingly, explicit representation of a more specific instance may be lost in this case. On the contrary, labeling a *mule deer* as a *deer* can otherwise verify the hypernym-hyponym relationship between the two subjects, and hence helps understanding that *deer* is a grouping of a more specific subclass *mule deer* By a continuous cycle of image validation and dictionary commentary, the precision of a mapping between an image and a synset will improve over time.

4.4.4. *Validating Images*

A synset/image association is established from a variety of sources : user uploads, user free association, PicNet guesses, or the initial automatic PicNet seeding. The quality of these associations may be determined by different users through a voting process. In such a voting system, a user is shown a synset/image association randomly drawn from PicNet, with the caveat that those associations which have already been rated by the user, and those which receive sufficiently negative ratings be excluded. The user, when shown with this synset, a short defining gloss, and picture, will be asked to comment using the following options : this image (a) is NOT related to this concept; (b) is loosely related to this concept; (c) is related to many attributes of this concept; (d) is loosely related to this concept; (e) is well related to this concept. The vote is then recorded, the user shown a new pair. Figure 4.1 shows a snapshot of the PicNet validation screen.



(noun) **fish**

'any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills; "the shark is a large fish"; "in the livingroom there was a tank of colorful fish"'

This image:

○ is NOT related to this concept.

○ is loosely related to this concept.

○ is related to many attributes of this concept.

○ is well related to this concept.

○ No Reply

Figure 4.1. Image validation screen in PicNet.

4.4.5. *Competitive Free Association*

A gaming process was also devised to motivate users while they compete with one another. A minimum of five players is required to start the game, and only so with a majority of votes to start i.e. at least three must agree to start for the game to commence. In each round, a player is shown a picture and invited to provide a name for it. Identical answers are grouped together. Next, all answers entered previously are shown and each player now chooses the best answer that is not his own. The answer which receives the most votes win this round, as is the person who enters that answer. When multiple users enter the winning word, points are divided equally among these players. Word/image associations are then integrated into PicNet database, with each association scoring on the number of votes it receives in the game. This game may be played at a comfortable pace over several days. Users can elect when they want to login and continue, all at their convenience. To keep the game moving forward when a player fails to take his turn, PicNet will spoof the human player, making an entry for him, based on a number of precursors such as scoring information, educated guessing, and Wordnet relations in the PicNet database. Based on initial evaluations of this gaming mechanism, a number of improvements are made. In one instance, players often fail to provide a precise name for a specific image shown, e.g. naming *bone* for *humerus*. To correct this problem, synset mappings (from earlier automatic seeding and/or image assignment by an expert user) from PicNet are now included as one of the options in the voting process, with the hope that users will realize the existence of a more precise mapping and vote for it.

4.5. Ensuring Data Quality

Images collected from Web-users promises free flow data at little or no cost. However, there is a concern with this type of data collection method, usually the quality of contributions. PicNet realizes this and implements two schemes to ensure the quality of images stays at a satisfactory level.

4.5.1. *Scoring Sysnet/Image Association*

PicNet maintains a complete history of users activities, which is now used to rank synset/image associations. For each action carried out by the user, there is an implicit quantified vote related to each such association. These votes are summed up and culminates in a score for the pair, giving PicNet the ability to dynamically rank images associated with a particular synset. The following lists each user activity and its corresponding votes.

+5 Upload an image for a selected synset (from Search results)
+4 Image Validation (is well-related to this concept)
+3 Image Validation (is related to many attributes of this concept)
+1 Image Validation (is loosely related to this concept)
-5 Image Validation (is NOT related to this concept)

+3 Free Association

+n Competitive Free Association (where n=number of players agreeing with the association)

4.5.2. *Administrative Functions*

Given that potential errors may be introduced to PicNet database, either accidentally or deliberately, there is a need for measures that serves to protect its integrity. Also, there are strict rules regarding obscene images on government websites. These problems can be solved by assigning administrator rights to a superuser who filters out user contributions based on their quality and appropriateness. Particularly, images that are too small to be useful, containing porn, copyright-restricted, etc are not added to the database. They are discarded by the superuser through a verification process. Such verification process can be automated especially when judging on the size and content of the images. It may seem that a verification process acts as a bottleneck, but it is necessary. Moreover, validating images is compelling to a curious mind and can be performed at a fast rate of 24 images per minute. To streamline the process, reliable and trustworthy users may be assigned such administrator rights in the future. Should there be any malicious attempts detected, the user would be blocked forever and his activities rolled back, made possible by PicNet history mechanism.

4.6. Preliminary Evaluations

Experiments were conducted to evaluate on two aspects of PicNet system. (1) Average concurrence among users voting on the appropriateness of an image using the *Competitive Free Association* activity. (2) Quality of the top ranked word/image associations, based on the score summed up from all user activities in PicNet. For the first evaluation, the concurrence was given by the number of users who voted for the same synset suggestion in each round. The average concurrence was 43% with a standard variance of 0.05, which indicates that three out of five users reach a common agreement in that round. For the second evaluation, a manual inspection of all top ranked word/image associations suggested a very good mapping between each synset and its image.

More details on the PicNet system, its evaluation, and examples of word/image associations created with this system are available in [6].

CHAPTER 5

UNDERSTANDING WITH PICTURES

5.1. Motivation

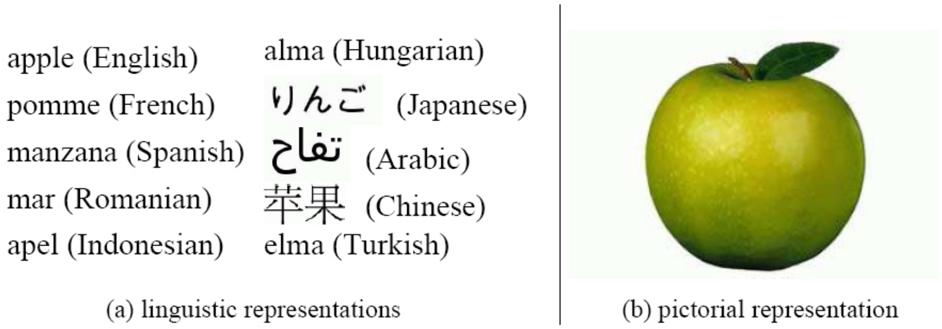
As explained earlier, there are more than 7,000 languages spoken worldwide, out of which only 15–20 languages can currently take advantage of the benefits provided by machine translation, and even for these languages, the automatically produced translations are not error free, and their quality lags behind the human expectations.

In this chapter[1], we investigate a new paradigm for translation: **translation through pictures**, as opposed to translation through words, as a means of producing universal representations of information, which can be effectively conveyed 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 would have a different way of referring to the concept of *apple*, as illustrated in Figure 5.1 (a). Instead, a picture would be understood by all people in the same way, thereby replacing the multitude of linguistic descriptions with one, virtually universal representation (Figure 5.1 (b)).

Such visual comprehension requires minimal active learning and in fact, is acquired since childhood through our everyday interactions with the outside world. Images continue to roll into our minds and become conceptually connected to the entities in our world knowledge, usually before we learn the linguistic representation of these entities. Hence, in addition to enabling communication across languages, the ability to encode information using pictorial representations has other potential benefits, such as language learning for children or for those who learn a foreign language, communication to and from preliterate or non-literate people, or language understanding for people with language disorders.

[1] Much of materials here and in chapter 6 are taken from two papers we submitted to CogSci 2006 [20] and AAAI 2006 [19]



(a) linguistic representations

(b) pictorial representation

FIGURE 5.1. Linguistic and visual representation for the concept "apple".

Previous work has focused merely on the understanding of pictures representing single concepts [25]. As outlined in our related work, their research focus mainly on understanding concrete nouns in the context of an all-words sentence, or the design of iconic languages for augmentative communication for people with speech impediments [8]. Instead, in this thesis, we evaluate the hypothesis that entire short sentences (e.g. *"The house has four bedroom and one kitchen."*) can be translated into pictures, which eventually could be understood independent of any language-specific representations. Figure 5.2(a) shows an example of the pictorial translations that we target.

There are of course limitations to this approach. First, there are complex informations that cannot be conveyed through pictures, e.g. *"An inhaled form of insulin won federal approval yesterday,"* which require the more advanced representations that can only be encoded in language. Furthermore, such understanding requires mastery of an advance level of vocabulary and a good grasp of the grammar. Second, there is a large number of concepts that have a level of abstraction that prohibits a visual representation, such as e.g. *politics*, *paradigm* or *regenerate*. Finally, cultural differences may result in varying levels of understanding for certain concepts. For instance, the prototypical image for *house* may be different in Asian countries as compared to countries in Europe. Similarly, the concept of *coffee* may be completely missing from the vocabulary of certain Latin American tribes, and therefore images representing this concept would be difficult to understand by the speakers of such 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 "translated through pictures," as compared to the more traditional linguistic translations. Note that we do not attempt to represent complex states or events (e.g. emotional states, temporal markers, change) or their attributes (adjectives, adverbs), nor do we attempt to communicate linguistic structure (e.g. complex noun phrases, prepositional attachments, lexical order, certainty, negation). Instead, we focus on generating pictorial translations for simple sentences, using visual representations for basic concrete nouns and verbs[2], and we evaluate the amount of understanding that can be achieved with these simple visual descriptions as compared to their linguistic alternatives.

Starting with a given short sentence, we use an electronic illustrated dictionary (PicNet) and state-of-the-art natural language processing tools (explained in Chapter 3) to create a pictorial translation. A number of users are then asked to produce an interpretation of these visual representation, which we then compare with the interpretation generated based on a linguistic representation of the same information. We provide the results later in the chapter.

5.2. Understanding with Pictures

The hypothesis guiding our study is that simple sentences can be conveyed via pictorial representations, with limited or no use of linguistic descriptions. While linguistic expressions are certainly irreplaceable when it comes to complex, abstract concepts such as *materialism*, *scholastics*, or *formalism*, simple concrete concepts such as *apple* or *drink* can be effectively described through pictures, and consequently create pictorial representations of information.

[2] Previous, similar research exclude verbs

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.

5.2.1. *Translation Scenarios*

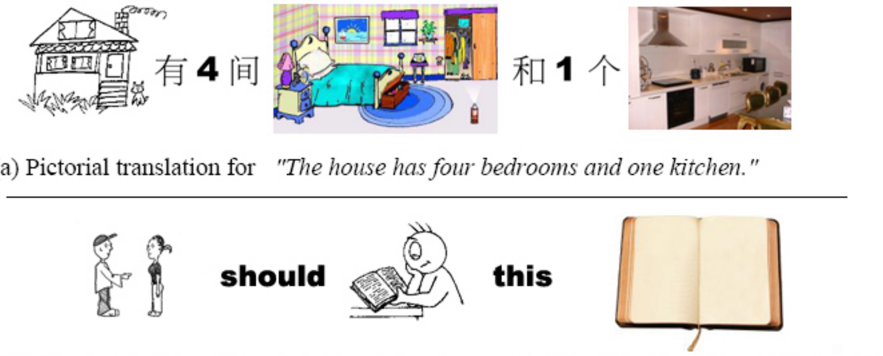
We conduct our experiments under the assumption that there is a **language barrier** between the two participants in an information communication process. The sender (speaker) attempts to communicate with a receiver (listener), but the only communication means available is a language known to the sender, 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[3]. 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 5.2(a) shows an example of such a pictorial translation.

[3] The task here is to translate from Chinese to English, and Chinese is an unknown language to the listener. Also, note that there is no clarification dialogue between the source and the target subjects

Scenario S2. An automatic language translation tool is available, which is coupled with a pictorial translation tool for a dual visual-linguistic representation. The linguistic representations in the target ("known") language are produced using an automatic translation system[4], and therefore not necessarily accurate. Figure 5.2(b) shows an example of a mixed pictorial-linguistic translation.

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 translations 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 5.2(c).



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

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

I eat the egg and the coffee work as breakfast.

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

FIGURE 5.2. Sample pictorial and linguistic translations for three input texts.

[4] We use SYSTRAN http://www.systransoft.com

5.3. Translation into Pictures

The pictures required for our experiments are collected from PicNet, described earlier in chapter 4. 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. A manual inspection of 100 random concept/image pairs suggests that the scoring scheme is successful in identifying high quality associations, with about 85 associations found correct by trusted human judges (Figure 5.3). In our picture translation experiments, we construct an automatic system for translating words into pictures, achieved using a few state-of-the-art natural language processing tools and coupled with PicNet. For continuity purposes, we will only discuss in details this automatic pictorial translation system later in chapter 6. Once again, no attempt is made to assign pictures to adjectives or adverbs. There are also many complex nouns and verbs that do not have a visual representation, in which case no pictorial translation is performed. In addition to the nouns and verbs found in PicNet, pronouns are also represented in the pictorial translations, using images from a language learning course[5].

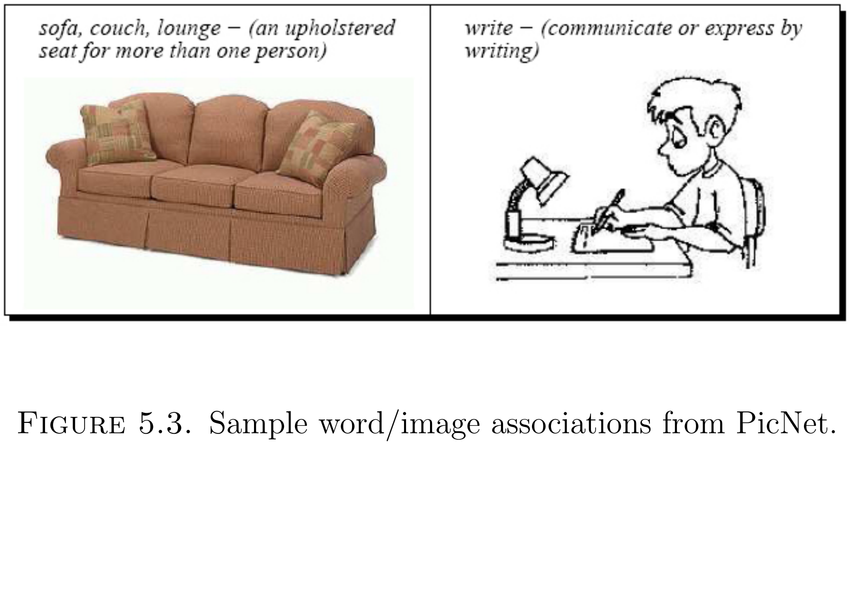

sofa, couch, lounge – (an upholstered seat for more than one person)
write – (communicate or express by writing)

Figure 5.3. Sample word/image associations from PicNet.

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

5.4. Experimental Setup

We first attempt to determine the amount of understanding that can be achieved using pictorial translations. We create a testbed of 50 short sentences, consisting of 30 randomly selected examples from language learning courses, and 20 sentences from various domain-specific texts[6], covering fields such as e.g. financial, sports, travel, etc. While all the sentences in our testbed are short, with an average of 10-15 words each, they have various levels of difficulty, ranging from simple basic vocabulary taught in beginner language classes, to more complex sentences containing domain-specific vocabulary. For each sentence, a Chinese translation is also available, created by three Chinese native speakers, which constitutes the "unknown" language for the translation evaluations. The 50 sentences (in the "known" language) are given in the appendix, with the first 30 from language learning courses, and the last 20 from SemCor[7].

Three representations are produced for each sentence: (1) A pictorial translation, where verbs, nouns, and pronouns are represented with pictures, while the remaining context is represented in Chinese (no pictorial translations are generated for those verbs or nouns that are not available in PicNet). (2) A mixed pictorial and linguistic translation, where verbs, nouns, and pronouns are still represented with pictures, but the context is represented in English. (3) A linguistic translation, as obtained from a machine translation system, which automatically translates the Chinese version of each sentence into English; no pictorial representations are used in this translation.

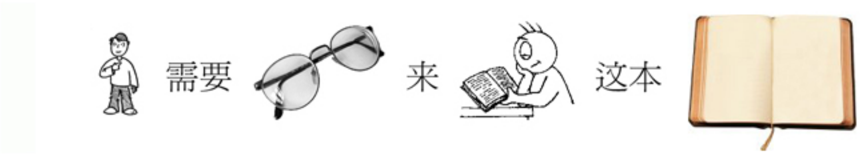
5.4.1. *Users Test on Interpretation*

Each of the three translations is then shown to a number of users, who have to indicate in their own words their interpretation of the visual and/or linguistic representations. For instance, Figure 5.4 shows a pictorial translation for the sentence *"I need glasses to read*

[6] SemCor, semantically annotated texts with WordNet 2.0 senses

[7] SemCor is a sense-tagged corpus with texts from different domains such as sports, travel, politics, financial etc

this book," and three interpretations by three different users[8]. Note that, to avoid any bias, the users will see the linguistic translations only after indicating their own interpretation of the pictorial representations. We conducted these surveys using paper-based and electronic-based methods. In the paper-based method, we engage users on a one-to-one basis to carry out the survey; though no time limit is set, we requested them to go by their first intuition and write down their interpretations in a prompt manner. The electronic-based method consists of a website which users can login and do the survey at their convenience. Results from both methods are collected.



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 5.4. Various interpretations by different users for a sample pictorial translation.

5.5. Experimental Results

To assess the quality of the interpretations generated by each of the three translation scenarios described before, we use both manual and automatic assessments of quality, based on metrics typically used in machine translation evaluations. These metrics are described in detail in Chapter 3.

For each sentence in our testbed and for each possible translation, we collected interpretations from fifteen different users[9]. No Chinese speakers were allowed to participate

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

[9] This is the average. Some translations for a few sentences have significantly more interpretations

in the evaluations, since Chinese was the “unknown” language used in our experiments. Moreover, the user group included different ethnic groups, e.g. Hispanics, Caucasians, Latin Americans, Indians, accounting for different cultural biases. 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 5.1 shows the evaluation results, averaged across all users and all sentences.

	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

Table 5.1. Results for the three translation scenarios, using automatic and manual evaluation criteria.

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% [9], which can be considered as a rough estimation of the upper bound.

5.6. Discussion

The results indicate that a significant amount of the information contained in simple sentences can be conveyed through pictorial translations. The human adequacy score of 3.81, also reflected in the automatic NIST and GTM scores, indicate that about 87%[10] of the information can be effectively communicated using pictures. This is probably 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.

The score achieved through the pictorial translations alone 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 basically 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) point out the role played by context that cannot be described with visual representations. Adjectives, adverbs, prepositions, abstract nouns and verbs, syntactic structure, etc. constitute a linguistic context that cannot be represented with pictures, and which nonetheless have an important role in the communication process.

Finally, the gap between the second and the third 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

[10] The fraction of the adequacy score for pictorial translations divided by the adequacy for linguistic translations.

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.

In another example, the presence of difficult concrete terms *hardtack* and *lobscouse* in *"He dumped the pan of crumbled hardtack into the boiling pot of lobscouse"* leads to more **generalized** interpretations such as *biscuits*, *cookies* and *pot of meat*, *meat* respectively. These evidence indicate that elaborate, unusual nouns are not suitable for translation into pictures, as most users lack the vocabulary to interpret them into linguistic terms, and can only generate conceptual understandings that lack specific details.

Interestingly, there were also cases where the interpretation of the pictorial translation 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*.

Pictorial translation also fares better over current state-of-the-art linguistic translation in cases where morphological differences between the source and target language causes a wrong translation that skews overall interpretation of the text. Oddly, one of the four morphemes used to make up the Chinese word for *Oklahoma* was translated to *Russia*, probably due to statistical correspondence between them. Hence the sentence *"There have been three tornadoes in Oklahoma"* was interpreted by many users to be *"There have been three* *tornadoes in Russia"*. This interpretation is considered as a serious flaw where geographical significance is important.

Overall, while pictorial translations have limitations in the amount of information they can convey, 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 translations can be used for the purpose of communicating simple pieces of information.

CHAPTER 6

AUTOMATIC TRANSLATION SYSTEM

6.1. Automatic Pictorial Translation

In this chapter, we present our automatic pictorial translation system that was used to generate the test sentences for our experiments. This system may have potential uses in other areas like producing pictorial sentences for story-telling to children or as a tool for second language learning. The automatic translation of an input text into pictures is a non-trivial task, since the goal is to generate pictorial representations that are highly correlated with the words in the source sentence, thus effecting a level of understanding for the pictorial translations which would be comparable to that for the linguistic representations alone. The system includes the following main operating procedures :

6.1.1. *Tokenization*

We use an effective English tokenizer [3] that splits a sentence into individual, meaningful tokens that become candidates for translating into pictures. Assuming data entry errors by the humans e.g. misplaced periods, apostrophes, commas etc, the tokenizer cleans up each token and prepares it for the next step. An instance would be separating the period at the end of "The house has four bedrooms and one kitchen." while retaining it in "Dr. Sanjay Gupta".

6.1.2. *Part-Of-Speech (POS) Tagging*

POS tagging is necessary to identify the syntactic class a token belongs to. We use the Penn Treebank Tagset for labeling our tokens and use Brill's POS tagger [7], a rule-based tagger that is fast and performs to a acceptable level of accuracy of 93-95%. The output of this step is prepared for the next step, which is to return us the "root" of a given word.

6.1.3. *Lemmatization*

The lemmatization process is concerned with the identification of the base form for each input word, e.g. identify "room" as base form for "rooms" or "be" as a base form for "was". As a very simple but quite powerful lemmatization tool, we use a function from the Wordnet interface (validForms() from WordNet::QueryModule [1]). The lemma would be the shortest valid form of a word retrieved by the function. This step is necessary for two reasons. First, to identify the sense of a word, we are required to identify its lemma. Secondly, synsets in PicNet are organized only according to the lemma of any given word. In this case, we want to avoid redundancy i.e. words like "houses", "house", "housing" are lemmatized to a single concept "house" in a synset.

6.1.4. *Word Sense Disambiguation*

We attempt to pick out the sense of each open-class word most likely intended by users of our automatic system using SenseLearner2.0 [18], 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. Given that we do not have prior knowledge of the words likely to be translated, we have to contend with a marginal error in reaching the correct sense. Put in other words, we should not expect that SenseLearner will pick out the correct sense every time.

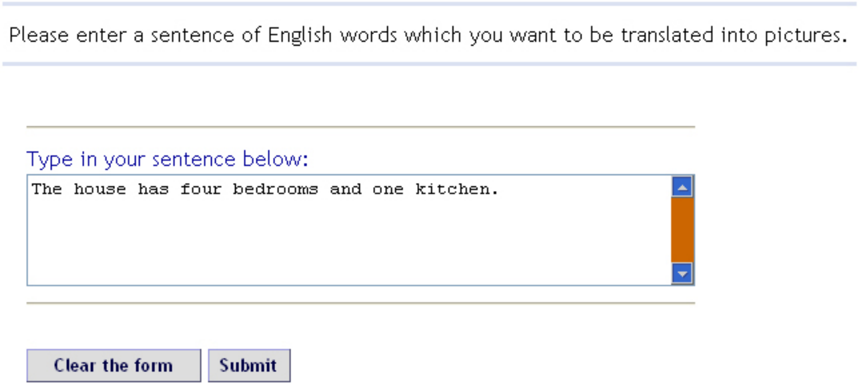
6.1.5. *Word-Picture Mapping using PicNet*

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 translation, we use

[1] http://search.cpan.org/ jrennie/WordNet-QueryData-1.39/

only those concept/image associations that rank above a threshold score of 4, indicating a high quality association.

Note that in addition to the image representations for basic nouns and verbs, as collected through PicNet, we also use the language learning course mentioned earlier to gather pictorial representations for pronouns, which are also used in the translations.

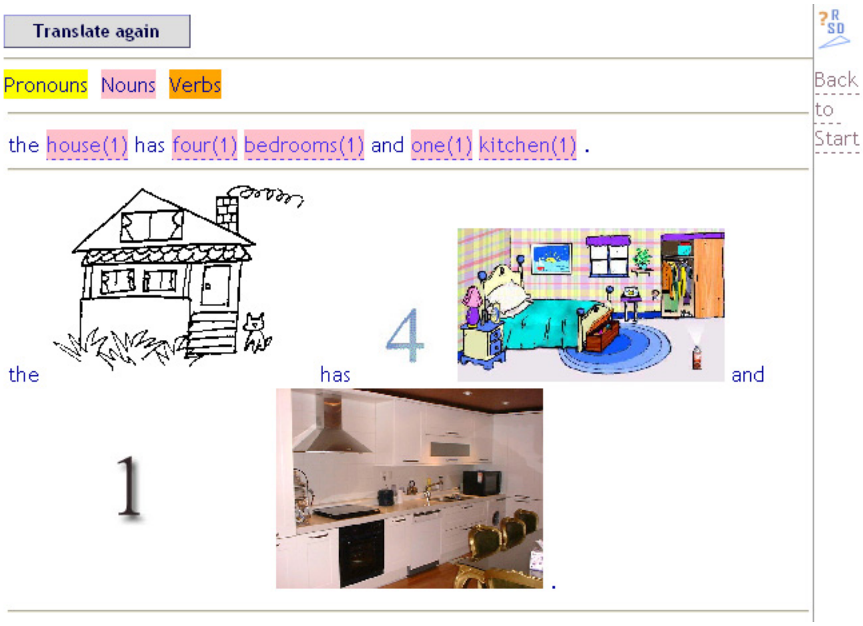


Please enter a sentence of English words which you want to be translated into pictures.

Type in your sentence below:

The house has four bedrooms and one kitchen.

FIGURE 6.1. Automatic translation system.



the house(1) has four(1) bedrooms(1) and one(1) kitchen(1) .

FIGURE 6.2. Translated pictorial representation.

6.2. Evaluation

Through our experiments, we target the evaluation of the translation quality for each of the three translation scenarios described before, as well as an evaluation of the system quality and usefulness, as determined by users of the system.

6.2.1. *System Quality*

In addition to the translation quality evaluation described in detail in chapter 5, we also conducted a user study concerning the quality and usefulness of the system. The study was designed using guidelines typically followed in human-computer interaction evaluations [14]. Twenty users (different than the users who participated in the translation evaluation) were asked to use the system to generate pictorial translations for 5–10 sentences of their own choosing. The users were then asked to use a scale from 1 to 10 to evaluate the system along four dimensions: (1) correctness; (2) interactivity; (3) intelligence; and (4) usefulness. Table 6.1 shows the user study questionnaire, and the average scores.

The system is producing correct word/image associations.	8.10
The system is interactive.	8.00
The system behaves intelligently.	7.21
Overall potential of the system to generate pictorial pictorial translations and its potential usefulness	7.84

Table 6.1. Evaluation of system quality.

The results indicate that overall the users were pleased with their interaction with the system, and found that the system produced most of the times correct word/image associations (8.1/10). The users also found the system to be interactive (8/10) and intelligent (7.2/10), and positively evaluated its potential usefulness (7.8/10).

CHAPTER 7

CONCLUSIONS

Language can sometimes be an impediment in communication. Whether we are talking about people who speak different languages, students who are trying to learn a new language, or people with language disorders, the understanding of linguistic representations in a given language require a certain amount of knowledge that not everybody has.

In this thesis, we described a system that can automatically generate pictorial representations for simple sentences, and proposed "translation through pictures" as a means for conveying simple pieces of information across language barriers. We carried out our experiments in three translation scenarios : fully pictorial representations, a mixture of pictorial and linguistic representations and fully linguistic representations. Through the results of our evaluation, we have shown that pictorial translations can generate a significant level of understanding. This is a great improvement over the baseline of zero communication when the source speaker and the target listener do not share a common language. Additionally, a mixed pictorial/linguistic representation pictorial with linguistic representations can generate understanding that is competitive with that of linguistic representations, suggesting that simple sentences with basic concrete nouns and simple verbs can indeed be communicated through sequences of pictures which replace those nouns and verbs. A user study was also conducted around the pictorial translation system and revealed that users found the system to generally produce correct word/image associations, and rated the system as interactive and intelligent.

Future work will consider the analysis of more complex sentences of various degrees of difficulty. The use of a larger scope of verbs may be possible. It is also likely to improve our pictorial representations using ontology methods, morphing sequences or collated pictures. Cultural differences in picture interpretation are also an interesting aspect that we plan to consider in future evaluations. This work may extend to involve other visual representations, such as video and/or audio, which augment pictures in our translations.

APPENDIX

EVALUATION SENTENCES

(1) This cat is four years old .

(2) He gives the child a dime .

(3) You can buy a used car at a low cost .

(4) Please sit down on this chair .

(5) I am taking a computer course at a local college .

(6) You should go to a doctor for that cold .

(7) Cotton is used to make clothes .

(8) I read his latest column in the New York Times .

(9) I eat an apple after dinner .

(10) The bank closes at three in the afternoon .

(11) He bought a new boat for his birthday .

(12) Can you get some bread from the supermarket ?

(13) My brother lives in Seattle .

(14) You should read this book .

(15) Will you like to go dancing with me this Saturday ?

(16) I visited my dad last week .

(17) Will you like the boiled egg or fried egg ?

(18) He milks the cow everyday .

(19) He drinks two glasses of water .

(20) I eat eggs and coffee for breakfast .

(21) I will travel to Africa .

(22) I bought a pair of new shoes last week .

(23) There have been three tornadoes in Oklahoma .

(24) I need my glasses to read this book .

(25) I wrote a letter to my mother .

(26) I read email on my computer .

(27) Please bring me a glass of tea and a biscuit .

(28) The house has four bedrooms and one kitchen .

(29) I go to the gym to exercise .

(30) I like to eat milk and cereal in the morning .

(31) He saw the sign above the door of the hut : Home Sweet Home .

(32) He dumped the pan of crumbled hardtack into the boiling pot of lobscouse .

(33) He settled on the sofa with his coffee, warming his hands on the cup .

(34) He found the pilot light and turned on one of the burners for her .

(35) The portable record player with a pile of classical records beside it .

(36) He reached out and felt the bath towel hanging on the towel rack over the tub .

(37) They took Jesus 's body, then, and wrapped it in winding-clothes with the spices .

(38) David reached for the pair of pistols in the saddlebags at his feet .

(39) The fish took the bait .

(40) He could see the bright torches lighting the riverbank .

(41) In the corner was the soldier with the white flag .

(42) She lay still on the bed, her head hardly denting the pillow .

(43) Her legs hung down long and thin as she sat on the high stool .

(44) He finally fell asleep around six in the morning with the aid of a sleeping pill .

(45) In one hand she clutched a hundred dollar bill and in the other a straw suitcase .

(46) That couple has a son and a daughter .

(47) Tanks lined up at the border will be no more helpful .

(48) The sick were always receiving medicines .

(49) The bottle was filled with flour .

(50) There was a lady there , in a pyjamas .

BIBLIOGRAPHY

[1] Barton A., Sevcik R., and Romski M., *Exploring visual-graphic symbol acquisition by pre-school age children with developmental and language delays*, Augmentative and Alternative Communication 22 (2006), 10–20.

[2] Way A. and N. Gough., *Comparing example-based and statistical machine translation*, Natural Language Engineering 11 (2005), 295–309.

[3] Y. Al-Onaizan and Dan. Melamed, *Statistical machine translation*, JHU Summer Workshop, 1999.

[4] K. Barnard and D.A. Forsyth, *Learning the semantics of words and pictures*, Proceedings of the IEEE International Conference on Computer Vision, 2001.

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

[6] A. Borman, R. Mihalcea, and P. Tarau, *Picnet: Augmenting semantic resources with pictorial representations*, Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (Stanford, CA), 2005.

[7] E. Brill, *A simple rule-based part of speech tagger*, Proceedings of the 3rd Conference on Applied Natural Language Processing (Trento, Italy), 1992.

[8] S. Chang and G. Polese, *A methodology and interactive environment for iconic language design*, Proceedings of the IEEE workshop on visual languages, 1992.

[9] C. Corley, A. Csomai, and R. Mihalcea, *Text semantic similarity, with applications*, Proceedings of the International Conference on Recent Advances in Natural Language Processing (Bulgaria), 2005.

[10] W. Daelemans, J. Zavrel, K. van der Sloot, and A. van den Bosch, *Timbl: Tilburg memory based learner, version 4.0, reference guide*, Tech. report, University of Antwerp, 2001.

[11] 2005, http://www.ethnologue.com.

[12] W.W. Gibbs, *Saving dying languages*, Scientific American (2002), 79–86.

[13] Pan Jia-Yu, Yang Hyung-Jeong, Christos Faloutsos, and Pinar Duygulu., *Gcap: Graph-based automatic image captioning*, Proceedings of the 4th International Workshop on Multimedia Data and Document

Engineering (MDDE 04), in conjunction with Computer Vision and Pattern Recognition Conference (CVPR 2004) (Washington DC), July 2004.

[14] H. Liu, H.; Lieberman and T. Selker, *A model of textual affect sensing using real-world knowledge*, Proceedings of the ACM Conference on Intelligent User Interfaces, 2003.

[15] Argyle M., *Bodily communication*, International Universities Press, 1990.

[16] Hauser M., Chomsky N., and Fitch W., *The faculty of language: what is it, who has it, and how did it evolve?*, Science 298 (2002), 1569–1579.

[17] D.I. Melamed, R. Green, and J. Turian, *Precision and recall of machine translation*, Proceedings of the HLT-NAACL 2003, Short Papers (Edmonton, Canada), 2003.

[18] R. Mihalcea and A. Csomai, *Senselearner: Word sense disambiguation for all words in unrestricted text*, Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (Ann Arbor, MI), 2005.

[19] R. Mihalcea and B. Leong, *Towards communicating simple sentences using pictorial representations*, Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-2006) (Submitted) (Boston), July 2006.

[20] R. Mihalcea and B. Leong, *Understanding with pictures*, Proceedings of the 27th Annual Conference of the Cognitive Science Society (CogSci 2006) (Submitted) (Vancouver), July 2006.

[21] G. Miller, *Wordnet: A lexical database*, Communication of the ACM 38 (1995), no. 11, 39–41.

[22] G. Miller, M. Chodorow, S. Landes, C. Leacock, and R. Thomas, *Using a semantic concordance for sense identification*, Proceedings of the 4th ARPA Human Language Technology Workshop, 1994, pp. 240–243.

[23] F. Och and H. Ney, *Improved statistical alignment models*, Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (Hongkong), October 2000.

[24] K. Papineni, S. Roukos, T. Ward, and W. Zhu, *Bleu: a method for automatic evaluation of machine translation*, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002) (Philadelphia, PA), July 2002.

[25] M.C. Potter, J.F. Kroll, B. Yachzel, E. Carpenter, and J. Sherman, *Pictures in sentences: understanding without words*, Journal of Experimental Psychology 115 (1986), no. 3, 281–294.

[26] 2006, http://en.wikipedia.org/wiki/Visual_language.