[From: *International Journal of Translation*, vol.13, no.1-2, Jan-Dec 2001]

Word Sense Disambiguation
through Visual Thought Analysis

SRINTVASAN VAIDYARAMAN

INTRODUCTION

The areas of Machine translation, Natural language searching and programming are branches of Natural language processing (NLP) and any real progress in these areas requires breakthroughs in NLP. There have been diverse approaches to tackle the various problems of natural language processing. But they have been handicapped by one or more serious drawbacks. Specifically, assignment of meaning to words based on context has proved to be a forbidding problem. To make headway in this difficulty, it seems a fresh and bohemian approach is required. The satisfaction of this objective resulted in our method theorizing the relationship between visual and linguistic thinking which finds application in areas such as Machine Translation, Searching of databases and knowledge bases, Natural language programming and Machine intelligence.

BACKGROUND

Interest in NLP dates back to the 1950's particularly in its application to Machine Translation. However, after a relatively brief period of research activity the interest especially in the U.S. sagged because of the negative report of a committee set up to evaluate the feasibility of machine translation. Research in core NLP continued and was exemplified in various software programs that tried to mimic human understanding of language. Due to a lack of application of a suitable hypothesis of human understanding of language, these systems were not fit for real-life projects.

With development of technology in other areas of computer science and also in areas other than computer science, the need for language understanding by machines became more relevant. For example search engines will become more accurate, effective and speedier if they can understand natural language. Similarly, automating replies to queries saves a considerable amount of work. Rapid strides in communication technology warrant a reduction in language barriers if it is to be commensurately effective. This becomes easy if machines can automatically translate from one language to another. Natural language programming is another area which is the next logical step in the field of Programming Languages and which requires language understanding by machines. Thus broadly, NLP is our area of focus and more specifically it is disambiguation of word sense in a sentence.

The problem of word sense disambiguation has been a hard one to tackle. The problem is, given multiple meanings of a word the machine has to select the correct meaning based on the context. Work in this area has been going on for a long time. We focus our attention on the currently widely used statistical techniques. These methods clarify the sense of the word with the help of statistics (such as the frequency) on the surrounding words. Multiple features may also be used. The technique is beset by a few drawbacks. Firstly, it is not simple and in an area in which its problem intrinsically becomes intricately complex with scaling up, a simple method would serve the purpose. Secondly, this doesn't seem to be the way the humans disambiguate sense. These methods don't conceptually rely on any hypothesis of human understanding of language. In our opinion, any system that is simulated needs to be understood well if the simulator is to faithfully reproduce the functionality of that system. The accuracy of statistical methods is also not up to the expected levels. A system that is 90% accurate is deemed to be satisfactory. However, because of the second drawback that we mentioned, we strongly believe that the goal is difficult to achieve. Our objective is to tackle the particular problem of word sense disambiguation that makes the machine interpret the meanings of a word in the supposed way humans interpret. The way our method deviates from the previous methods is to effectively use a hypothesis on the phenomenon of visual thinking and applying it to the problem of word sense disambiguation.

This makes our method intuitive and very simple to grasp. The implementation becomes a lot simpler and the system would scale up well. As we mentioned, our method is not arbitrary and is backed by theory. The system has been manually tested for a number of ambiguous word usages and based on it we expect the accuracy to be 90% or above for a full-scale system.

VISUAL AND VERBAL THINKING

It is our hypothesis that linguistic thinking is not an independent phenomenon but only a façade that masks the real mode of thinking viz. visual thinking. Although the hypothesis that we think in terms of images is not new and has been put forth amongst others by philosophers, the application of this idea to practical problems such as NLP is novel. The relationship between words and the images evoked by these words underpin our methodology. For each word or a sequence of words we associate an image or a series of images. We stipulate that information content of these images can be described in terms of "variables".

We seem to gather the following information content from the images. The associated variables are given.

1. Information about the presence of something. For example when we hear the word "dog", we visualize a dog. We indicate this by a variable called "presence".
2. Information about the position of something. When we hear the word "kick", we visualize the leg moving (i.e. changing position). When we hear the word "height", we visualize the distance between the top position and bottom position of the object. Thus the variable for this is called "position".
3. Information about speed of movement of an object. The variable is "speed".
4. Information about size of an object. The variable is called "size".

5. Information about shape of something. We call the variable as “shape”.
6. Temporal information of an event. We assume this is expressed either symbolically or through actual images. The variable for this category is called “time”.
7. Information on the extent of presence or occurrence of something. For example the words, “partly”, “deficiency”, “certain” etc. are of this category. When we hear the words “partly right”, we see the images that represent this situation. We call the variable as “extent”.
8. Information about the quantity of something. We call the variable for this as “Number”.

(Note: We may add color, brightness etc also as variables but we are ignoring it for brevity and only a few words need them in their definitions).

In all these cases, the images visualized depend on the experiences, culture etc of the person and may thus vary from person to person. Besides, mostly the visualization of the images is fast and sub-conscious and the person does not fully experience it. However there are times when images depicting words emerge to the realm of consciousness, for example, when the words contain a high degree of emotional content.

In addition to the above variables that can be visualized we have the following entries:

1. Speech
2. Thought
3. Emotion and Perception such as HAPPINESS, SADNESS, ANGER, SIGHT, HEARING, SMELL, FEELING etc.

These form a separate category because they are not usually visualized. For example the feeling of happiness cannot be visualized though the events that lead to happiness can be visualized. Though words indicating thoughts don’t make us visualize thoughts these words conjure up a particular image, for example a person with his head facing downwards.

(Note: Speech can actually be "sensualised" by perception (hear) in our mind and can be visualized, in a sense, indicated by the movement of lips. We call the above entries pseudo-variables.)

Thus by and large images depict and contain all the information which may be later conveyed as words. This information is represented by variables and pseudo-variables and this idea forms the crux of our method. Applying this concept of mapping the word content to surrogate image content which are variables and pseudo-variables, to our specific problem of discriminating among the meanings of a word, we proceed towards finding a satisfactory solution to the problem. Thus given a text, it can be represented by equivalent variables and pseudo-variables by replacing each word in the text with the variables and pseudo-variables. For example the word "handsome" may be replaced with perception (SEE) and emotion (HAPPY). At this point we digress a little as it is instructive to consider our following hypothesis on human understanding of language (We do not claim novelty of this hypothesis as it may have been put forward in various guises at earlier points of time. However where we claim novelty is in deriving a method for resolving ambiguity of a word sense that simulates human beings' own disambiguation of it.) We discuss our algorithm once we explain the hypothesis on human understanding of language.

WHAT ARE THOUGHTS?

It is our hypothesis that sensory perceptions form the basis of thoughts. We would like to separate thoughts into two distinct types. The first type occurs when we are directing our senses on to the physical world. For any biological system that has to respond to external stimuli from the physical world in an adaptive manner, memory plays an indispensable role. In a fundamental sense understanding can be considered as a response to the sensory stimuli. The manifestation of this response in the case of humans culminates in what we term as "recognition" being the evoking of mental imagery from memory. This act of recognition of an entity occurs when there is a match between what the system senses i.e., the entity and what it has already sensed i.e., its memory.

More generally, humans have the facility for approximate matching. So, in addition to specific objects being recognized, for example the face of the mother, any object of that type is also recognized, say any face. This is also referred to as detection. But we won't make this distinction and consider both as recognition. Similarly, movement of objects, which we call actions, is also recognized. Sensory perceptions are thus made up of a stream of recognitions which are labeled as thoughts. Without the recognition, sensory perceptions are merely pre-designed responses to the external stimuli which one does not become aware of.

In this first type we do not include the thoughts that occur as a result of communication through a language. We include only thoughts that occur as a result of perception. For example the following recognitions of a perceiver may fit this category: A big hotel in the front. A car going past. Rustling of the trees. A girl wearing a blue dress crossing the road. Fragrance of rose, etc. These events do not involve any interaction through language but only recognition of objects and actions. We also include in this category only those perceptions that are occurring in the present tense.

The second type of thoughts occurs when we are not directing our focus on to the external world or when that focus is at a reduced level. Our memory and the ability to manipulate the contents of the memory form the basis of these thoughts. In contrast to type 1 thoughts where the object of awareness is what that is being sensed, type 2 thoughts have their basis of awareness as what is in memory. Type 2 thoughts can be classified into two. In one, imagery of what is in memory is elicited and in the other, new images are created out of the images already in memory. In either case it is enough to know that recognition emerges as in the case of type 1 thoughts. Please note that we don't include thoughts that occur when we are not in a “conscious” state such as in dreams. Type 1 and Type 2 thoughts will suffice for our discussion of language understanding.

WHAT IS LANGUAGE?

Language according to me, in the case of type 1 thoughts, in a fundamental sense is the labeling in a particular order of the recognitions that make up the sensory perceptions. In the case of type 2 thoughts, language is the labeling in a particular order of the recognitions elicited from memory after search. Thus language is the expression of thoughts in an ordered manner.

It is now helpful to examine how the thought formation process occurs in the case of understanding. We start our discussion from how children learn language. Children initially learn the meaning of 'words through visual perception. It may be either through gestures or actual enacting of the meaning of the words. The meaning of the words thus being explained, their images are stored in the memory. When one of the learnt words is heard the corresponding stored images are evoked. New words are learnt from the combination of already learnt words or as explained before.

Images are stored into meaningful units, a unit being an object, or object-action combination. A word may produce one unit of image or more units of images stored as a sequence. In the case of objects typically one unit seems to be stored. Actions seem to be associated with object(s) when they are learnt. For example “walking” may be stored as “mother is walking” i.e., the image of mother walking is stored. This is because during our initial period of learning of words we are imparted the meaning of an action by associating with it an object. Thus as actions are learnt, sentences are simultaneously understood. As meaning of a word becomes complex more image units are stored for it.

It is our conjecture that as more instances of an action are heard, new images are formed from the previous images with the object being replaced. To illustrate this, let us consider that the child stores the image of the word “laugh” with its mother doing that action. In the memory of the child the image of its mother laughing is stored. Subsequently when the child hears the sentence “Father is laughing”, the image of “mother is laughing” materializes from its memory and the mother object is replaced with the father object. This constitutes the understanding of the sentence.

We do not want to go into the details of how the replacement of object takes place but it is clear we have that facility. For example, we might have seen a person kicking a ball. However when we hear the sentence, “He kicked the bat”, we are able to visualize it by replacing the ball with the bat and hence understand the statement. Image units seem to be stored together as a bundle with their length corresponding to the one that can be held by some sort of working memory. We might have experienced this when we read a long sentence. We are not immediately able to comprehend the whole sentence unless we go through it slowly. When we read it slowly, it gives time for the images to be formed and understood and stored and images of subsequent words formed and so on.

Let us take a longer sentence and analyze how the understanding emerges.

Consider the sentence, “Ram was playing the guitar while I was singing on the dais”. Consider the hypothetical situation where the working memory holds only two units of image. In this case the word “Ram” evokes images of “Ram” either with the strongest emotional content or the latest one (We leave this question open) and this process is not conscious. When the words “was playing” are added all the images that contain a person doing the action of “playing” something are evoked and the image of the person is replaced with that of Ram.

Assume that the best match of the images already stored in memory is “Shekhar is playing the violin”. Thus Shekhar is replaced by Ram. When the word “guitar” is added, violin is replaced by guitar and the image of “Ram was playing the guitar” is formed. Similarly “I” invokes the image of the speaker and the next sequence. of words “was singing on” elicits all the images of a person doing that action in a particular place. Please note that the meaning of the prepositions is associated with other word types say the verbs and nouns. For example in the above case separate images are not formed for “singing” and “on” but a single image is formed for “singing on”. Thus when these words are heard the place of action is also visualized. We assumed that the working memory can hold two units of images at a time and thus “Ram was playing the guitar” may be stored as a bundle and “I was singing on the dais” as another bundle. “While” may be indicated by the superposition of the two bundles.

Tense is understood using similar mechanisms i.e., by tagging of images corresponding to the various tenses. It is also likely that past tense evokes more image units than present tense as it implies an action that has been completed.

Understanding of language thus proceeds by evoking bundles of memory and pruning them by selecting the most closely matched bundle(s) as words are added and by replacing the objects already stored in the bundle(s) of images with the objects that are evoked when the sentence is uttered.

Relevance to contextual meaning in Machine translation and Information Retrieval

Now consider the sentence: “I went to school”. Assume the sentence is uttered up to “to” i.e., “I went to”. Here we are focusing on extracting the meaning of the word “school”. The part sentence may invoke the bundles “John went to museum” and “Mary is going to the church” based on the verb “went”. The evoked bundles have something in similar and also differ by the uniqueness of the words that make them up. Conceptually, the uniqueness of a word can be extracted by replacing that word in the sentence by another one and noting the difference in the image(s) that the new sentence creates. The similarities in the evoked bundles accommodates the semantics of the incoming word which is “school” but at the same time the uniqueness of the images of the words in a sentence generally enables its interpretation in a clear-cut manner.

Our algorithm is based on what we now explicitly describe. The complete procedure is as follows: We list the variables and pseudo-variables that make up the definition of the preceding and following word. We then list the variables and pseudo-variables that may precede and follow a particular word. For each meaning of our word ,a match is now performed between the variables and pseudo-variables that can potentially precede and follow our word and the definition (made of variables and pseudo-variables) of the following and preceding word. The meaning which produces the maximum number of matches is the contextual meaning. This is the basic idea. There are enhancements to this which can make it more effective.

By analogy to our hypothesized human understanding of contextual meaning, the definition of the preceding and following word is similar to the uniqueness of the words. Just as the image of the word with multiple meanings is changed by replacing the preceding or following word, the meaning of a word with multiple meanings is given by the definition of the preceding or following word. The similarities between evoked bundles is equivalent to the list of variables and pseudo-variables that may precede or follow our word.

To put it in a simple manner, we ask how do we understand the meaning of a word? It is based on its context. How do we know the context? It is by and large from the preceding and following word ignoring certain classes of words such as articles etc. It has been suggested that a small window of words is sufficient to clarify the sense of a word.) There is an interplay between the images of the sequence of words that permit a meaning and preclude others. The meaning unfolds as words are added. More elaborately, the variables and pseudo-variables compose the image of each word and their permissible unification based on one's experience and reasoning give sense to a sentence. It is very closely to this way our method works (we clarify shortly what we mean by “very closely”), matching the variables that can precede or follow a word and the variables that make up the preceding and following word.

We now illustrate the concept. Consider word A. Let the variable(s) that conies before it be emotion (HAPPY) and the variable(s) that comes after it be Position for the first meaning of word A. Let the variable that comes before it be Thought and the variable that comes after it be Perception (SEE) for the second meaning of word A. Let word B precede word A and word C follow word A. Word B is defined by Thought and word C id defined by Speed. Now there is a match between the potential variable that can come before the second meaning of word A and the variable composing the definition of word B. The variable is Thought. There is no other match. Thus the contextual meaning of word A is its second meaning since it produced the maximum number of matches.

A more concrete example is as follows: Consider the verb "support". Let us take the following two definitions.

1) bear the weight of
2) help.

Variables that precede or follow the first definition: size, speed extent, presence (thing). Variables that precede or follow the second definition: extent, position (change), quantity, presence (human).

Now consider the sentence: "The bridge supports heavy lorries"

Preceding word: Bridge. Definition of Bridge - presence (thing)
Following word: Heavy. Definition of Heavy - size, speed

The contextual meaning is (1) as there are three matches which are presence (thing), size and speed.

Knowing whether a variable comes before or after a word is important because order seems to play a vital role in human understanding of language. Consider the sentences: "The man saw the hill top" and "The top man saw the hill". The same words are used in both the sentences but the meanings of the sentences in general and the meaning of "top" in particular are different because of different ordering. Humans are able to disambiguate the meanings, which also suggests that there is unlikely to be a universal representation of sentences in terms of images. The image or image sequence is conditioned by how we learn language. Theoretically, by listing all the meaningful permutations of the variables, we may disambiguate the sense of the words (as is likely to be done by humans) but since the number of permutations would be very high, the approach that we suggested would be a good approximation for real-life situations since the window of words that dictate the sense of a word is very small. Also, our approach takes into consideration all the possible variables that may come before or after the word in question and hence there will automatically be a match if any of these variables precede or follow.

Illustration of problems in an application area (automatic translating systems)

We list below a few of the errors of the present translating systems. The first sentence is in source language. The second sentence is again in source language translated back from the target language by the system. The entry in bold letters show wrong meaning has been assigned. According to us there is enough context in these sentences that can be used to pick up the right meaning and which our method can detect. Nevertheless the current translating systems fail to do that.

la) He offered me handsome salary.
1b) It offered me beautiful salary.

2a) He told me a tall story.
2b) It told me a large history.

3a) He went through a lean period.
3b) It passed by one period thin.

This is a sample of meanings wrongly assigned by the system. The potential number of errors is obviously large. This is the case for relatively closely rooted languages such as English and French. The situation is worse for not closely related languages.

CONCLUSION

The problem of word sense disambiguation in the area of NLP was considered. The application of a hypothesis on human understanding of language was put forth as a solution. Specifically, the application involved representing the image content through what we termed "variables" and manipulating them to give sense to a word. In a sense these variables are akin to abstract categories but more importantly these categories have a firm backing and are ones that can be verified by anyone analyzing his thought process. We have tested some of the most ambiguous words for disambiguating their sense in a sentence and have found a high degree of accuracy. We expect the system to scale up well and perform with 90% + accuracy.