Integration of Background Knowledge in Language Processing: A Unified Theory of Metaphor Understanding, Moses Illusions and Text Memory

Raluca Budiu (Raluca.Budiu@cs.cmu.edu)
John R. Anderson (ja+@cmu.edu)
Department of Computer Science
Carnegie Mellon University, Pittsburgh, PA 15213

Abstract

One of the challenges of cognitive psychology is developing general models that explain a wide range of empirical phenomena. We describe a unique language comprehension model that fits data from several text comprehension domains: metaphor understanding, processing of semantic illusions and text memory. We show how background knowledge plays a similar role in all these processes, helping or hampering them. The model assumes that sentence processing at the semantic level is incremental, nondeterministic and incomplete, and uses background knowledge hints at each step. The model is implemented in the ACT-R framework (Anderson & Lebiere, 1998). The empirical phenomena that we model are: position effects on metaphor understanding, influence of distortion "quality" on semantic illusions and memory for related stories.

Introduction

While the role of background knowledge in language processing is widely acknowledged, the literature in the field has been less concerned with how background knowledge exactly contributes to building a text interpretation (with few exceptions — e.g. Sanford & Garrod, 1998).

In this paper, we will attempt to address the role that background knowledge plays in language processing. Namely, we propose a model which, at each moment, uses all the information available in the sentence to find a schema or script appropriate to the current situation; once identified, the content of that schema helps comprehension or recall. This view is in contrast with those models of comprehension which assume that the integration of the sentence with its context and/or background knowledge is done after the entire sentence has been read and the meanings of the individual words have been processed.

The model that we propose has the following qualities: (a) is **incremental**: after each word is input, the model searches the background knowledge for a sentence interpretation; (b) the model follows a **trial-and-error paradigm**: at each step a candidate interpretation is formed; if subsequently it proves wrong, it is rejected; (c) the model mixes **bottom-up with top-down strategies**: word meanings help finding

a sentence interpretation and the sentence interpretation modulates the meaning extraction processing; (d) the model is based on **incomplete processing at the word meaning level**: the number of word features processed may vary; (e) the model is based on **incomplete processing at the sentence level**: some of the information previously processed may be actually "forgotten" later on and not contribute to the sentence interpretation; (f) the model takes **thematic roles of words as input** and does not address syntactic processes.

This theory described in this paper has been implemented in the ACT-R framework (Anderson & Lebiere, 1998). ACT-R is a production-system cognitive architecture, which has been widely used to model a variety of problem-solving and memory tasks. The reason we chose this architecture is to show that there is nothing special about language processing: it can be implemented in this system in the same way as many other problem-solving tasks. An advantage of the ACT-R theory is that it allows models to produce concrete predictions (reading times, accuracies) that can be compared with the empirical data.

One of the major challenges of modeling sentence comprehension is that people are relatively fast at understanding sentences. The reading time for a typical sentence is on the order of a few seconds. While complicated processes may occur in sentence comprehension, they must happen in this short interval. The ACT-R implementation of our theory produces sentence processing times comparable to the times that people take in similar tasks. We have tested it against data pertaining to metaphor understanding and semantic illusions and text memory.

In what follows, we will first describe the basic sentence processing model; we will then continue with presenting the three sets of data to which the model has been applied: metaphor understanding, semantic illusions and text memory. Due to space constraints, we will focus more on the description of the general model and less on how the model performs the particular tasks that we are modeling.

The Sentence Processing Model

In this section we will describe how our sentence comprehension model works. We will start with presenting the skeleton of its behavior; on this skeleton we will incrementally add the more complex aspects of the model.

Our model asserts that understanding an isolated sentence involves finding a structure in memory that matches that sentence. This is equivalent to finding a background knowledge interpretation for the current sentence. Namely, when people are asked how many animals of each kind Noah took on the ark, presumably they identify the proper Noah story in their background knowledge repository and give an answer based on that finding. Also, when a metaphor like *Some jobs are jails* is read, people must access some general knowledge pertaining to jobs in order to understand it.

The representation for a background knowledge proposition is very simple and is exemplified in the Figure 1. As one can see from Figure 1, the background knowledge fact "Noah took the animals on the ark" is represented as a chunk in the long term memory (the node *Ark story* in the graph). This chunk is connected (via labeled links, which can be regarded as chunks themselves) to other chunks that represent the concepts involved in this proposition. After a sentence has been comprehended,

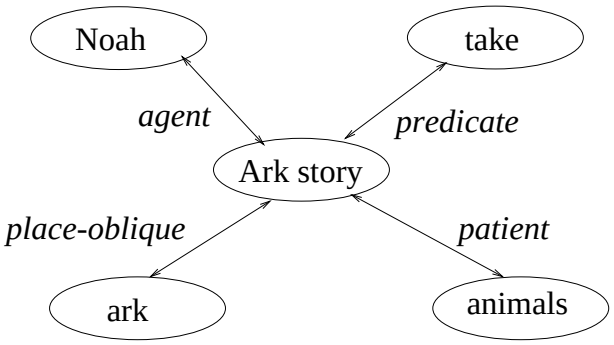

Figure 1: Representation for the background knowledge proposition *Noah took the animals on the ark.*

its representation in memory will be similar to the representation of this background knowledge proposition.

The model proposed is "eager" as opposed to "lazy": instead of postponing the process of finding an interpretation to the end of the sentence, the model searches for an interpretation as each word is read. A simplification that we make is to assume that for each word the model knows its correct thematic role, i.e. if it is an agent, a patient etc. After a new word is read, the model extracts its meaning from long term memory. Next, the model attempts to make sure that the current candidate context is consistent with the word being processed[1]. If there is no candidate context or if the current candidate context does not match the meaning, a new candidate that satisfies this constraint is sought for. When such a candidate is found, the model must make sure that it is consistent with all the meanings that have been processed previously. However, if the sentence is long enough, the model may "forget" to check all the previous words, and thus, the context may be constrained only by a few of them. The essential function of "forgetting" is to make the comprehension more forgiving. One can conceive the amount of "forgetting" as an individual-dependent parameter.

To give an example, suppose the model must process the sentence *How many animals did Noah take on the ark?* After reading *animals* a candidate context might be *My father raises ten animals on his farm.* Notice that the current word must have the same thematic role in the new sentence as in the background knowledge proposition that is considered as a candidate context, i.e. in this case *animals* should be a patient in the candidate context, since it is so in the actual sentence. If the next word read is *Noah*, the current candidate context is no longer viable, and a new one, matching *animals* as patient and *Noah* as agent, should be sought for. The process repeats until all the words in the sentence have been read.

Once sufficient words from the sentence have been processed, the model achieves a high degree of confidence that it has the right context. Then the context itself participates in the word comprehension process by priming the meaning of the word to be taken into consideration. The main difference between the model with context priming and the model without context priming is that, after reading the word, instead of extracting first the word meaning and then matching it onto the candidate context

[1] A context is consistent with a word with thematic role θ if the concept that plays the same role θ in the context is highly similar with the meaning of the word.

(if any), the corresponding role extant in the context is extracted and match onto the word meaning. In this way, the context-supported meaning of the word will be always extracted.

In an elaboration to this model, we have assumed that the meaning is a distributed set of features and we have made the meaning feature replace the word as the unit for context matching. In this view, processing a word meaning is equivalent with extracting some of its features. The model remains roughly the same. Suppose that the sentence to be processed is *Noah took the animals on the ark*, and that the word *Noah* had the features *patriarch*, *navigator*, *biblical-character* and *married*. Then, one might first extract the *biblical-character* feature and consider a candidate context involving a biblical character as an agent (for instance, a proposition about Jesus); after processing the *patriarch* feature, the current candidate context is rejected and a new one, involving an agent who is both a patriarch and a biblical character is looked for. It is not necessary that all the features of the word *Noah* be actually extracted; the number of features taken into consideration may depend on individual characteristics and on time pressure.

In the next sections, we will show how this model applies to metaphor understanding, Moses illusions and memory for text.

Metaphor Understanding

The literature on metaphor understanding is rich and controversial. A lot of attention has been paid to the question of processing seriality for non-literal sentences. Searle's serial position (Searle, 1979) states that literal comprehension is mandatory and precedes metaphorical comprehension. Only after failing to find a literal interpretation do subjects proceed with searching a metaphorical one.

Often this position has not been supported by empirical data. It has been repeatedly shown that in many situations, albeit not all, metaphor comprehension is as fast as literal comprehension(Ortony, Schallert, Reynolds, & Antos, 1978). An interesting question in this context is: are there metaphors easy-to-understand and metaphors hard-to-understand[2], and if so, what makes a metaphor harder? Do people select their comprehension strategy according to some perception of metaphor difficulty?

A good starting point in exploring this question is a study by Gerrig and Healy (1983). They found that the position of a metaphor in a sentence affects the reading time of the sentence. Thus, in their experiment, sentences with metaphors at the beginning (like *Drops of molten silver filled the night sky*) took significantly longer to read than sentences in which the metaphors occur at the end (*The night sky was filled with drops of molten silver*), and this result was not an artifact of the sentence structure. The actual reading times are given in Table 1.

The model nicely captures this effect. The metaphor-second sentences benefit from the existence of a strong context to guide the meaning processing. Thus, before reading *drops of molten silver*, a strong candidate context would be *The night sky was filled with stars*, which expects *stars* to be the theme of this sentence, and therefore enables processing of those features of *drops of molten silver* that are common with *stars*. However, after *Drops of molten silver filled* is read, the candidate context expects a

[2]Little attention has been paid to poetic metaphors. Metaphors typically used in psychological studies are ones that could be met in everyday speech.

Table 1: Mean reading times for metaphorical sentences — data (from Gerrig & Healy, 1982) and model.

Type of sentence	Reading Times (sec)	Model (sec)
Metaphor-first	4.21	4.23
Metaphor-second	3.53	2.84

container in the patient role; *sky* does not have many common features with a container, so context priming does not help. Moreover, the current candidate context must be abandoned in favor of a new one, corresponding to the correct interpretation of the sentence. This backtracking takes extra time, and, hence, the model predicts that metaphor-first sentences take longer to comprehend than metaphor-second sentences. As can be seen in Table 1, the ACT-R version of this model correctly reproduces the trends in the data.

Semantic Illusions

Erickson and Mattson (1981) were the first to draw the attention to the so-called Moses (or semantic) illusion. Semantic illusions refer to people failing to notice distortions in sentences like *How many animals of each kind did Moses take on the ark?*. When confronted with such a question, subjects usually respond *two*, even if they are asked to monitor for distortions and even if they know that it was Noah who took the animals on the ark.

Semantic illusions are a very robust phenomenon: most manipulations intended to make people more aware of the illusion failed (e.g. Reder & Kusbit, 1991). However, not all distortions are hard to notice. If *Moses* is replaced with *Adam* in the ark question, people are much less likely to fall for the illusion (Ayers, Reder, & Anderson, 1996). Also, certain "illusions" simply don't work: the illusion rate[3] for *Who was the first man to walk on the sun?* is zero.

Ayers et al. (1996) looked at illusion rates for good and bad distortions. They showed subjects three variants of the same distorted question, one containing a good distortion (*How many animals did Moses take on the ark?*), another containing a bad distortion (*How many animals did Adam take on the ark?*) and the third being the undistorted question (*How many animals did Noah take on the ark?*). As expected, subjects fell more often for the good distortions and less often for the bad distortions. They also showed a small bias to call a sentence "distorted" (for undistorted questions, they sometimes gave the "distorted" answer). The illusion rate column for the data in Table 2 presents these results.

From the fourth column in Table 2, one can see that the ACT-R model (based on the description in section) correctly captures these results. The model assumes that there is a meaning overlap between *Noah* in the ark story context and *Moses*, and also between *Noah* and *Adam*, but the number of salient features that are common to the

[3]Illusion rate is computed as the proportion of times people fell for the illusion, but knew the fact associated to the question.

Table 2: Illusion rates in the literal task and percent correct in gist task – data (from Ayers et al., 1996) and model.

	Illusion Rate (Literal)		%Correct (Gist)	
	Data	Model	Data	Model
Undistorted	.07	.08	82	86
Good Distortions	.46	.50	76	83
Bad Distortions	.29	.25	74	75

pair normal – good distortion is greater than the number of salient features common to the pair normal – bad distortion.

The model considers that a sentence is distorted if it is not able to produce an interpretation for it. Thus, when processing a sentence that contains a good distortion (*Moses*), the chance of extracting a feature that does not match the knowledge about the actual ark story is small, since the meaning overlap between the real agent of the ark story (*Noah*) and the distortion (*Moses*) is high. Hence the model will likely ignore such a feature. Moreover, even if a distinguishing feature is actually processed, it is possible that it be "forgotten" as the model tests a new context candidate.

On the other hand, if the feature overlap between the agent of the ark story and the bad distortion (*Adam*) is small, the chance of stumbling over a distinguishing feature is higher. Also, since these features are in a greater number, the probability of "forgetting" all of them in the subsequent candidate search is small.

An interesting variation of the Moses illusion task is the so-called gist task, in which people are asked to ignore the distortions, even if they notice them, and to try to give the "correct" answer to the intended question. Unlike for the "literal" Moses illusion task, people are very good at the gist task (Reder & Kusbit, 1991; Ayers et al., 1996). The column for the percentage of correct answers in the gist task in Table 2 shows some data adapted from the same study by Ayers et al. (1996). It is possible to make the model understand more distorted questions by making it more insistent in looking for an interpretation. Thus, if the end of sentence is reached with no interpretation, instead of deeming the sentence as uninterpretable, the model can continue the search for an interpretation, possibly by dropping features that have been unsuccessful before in helping to find a context. The model is actually very good at performing the gist task; in order to capture the less than perfect performance of subjects we had to assume a small bias of giving the "distorted" answer (see the last column of Table 2) .

While the model exhibits the behavior of the subjects in this task, it makes one fundamental assumption: that what distinguishes a good distortion (*Moses*) of a bad distortion (*Adam*) is that the former has more features in common with the corresponding role filler in the context (i.e. with *Noah* in the ark story). Oostendorp and Mul (1990), Oostendorp and Kok (1990) show that such a presumption is a realistic one. In these studies, they collected ratings of similarity between the distortion and the context in which it was supposed to appear (i.e. between *Moses* and the ark story), and they showed that subjects were more likely to fall for those distortions with high context similarity.

Table 3: Rate of recall per script version — data (adapted from Bower et al., 1979) and model. Number of actions is shown in parentheses.

Number of script versions	Data		Model	
	Stated actions	Unstated actions	Stated actions	Unstated actions
1	.38 (3.03)	.07 (.8)	.35 (2.80)	.05 (.60)
2	.28 (2.27)	.11 (1.26)	.37 (2.96)	.09 (1.08)

Script Effects

Since Bartlett (1932), a lot of studies have shown the influence of prior schemas on text memory. For instance, Bransford and Johnson (1972) showed that memory for text is improved if a setting is explicitly mentioned, allowing the readers to use their old knowledge about the topic in order to understand the text. Also, Owens, Bower, and Black (1979) found that when subjects were shown a series of distinct episodes (e.g. in the kitchen, at the doctor's, at supermarket etc.) linked together by a setting mentioned before the first episode (e.g. referring to a pregnant college student), the setting modulated the recall. In a related study, Bower, Black, and Turner (1979) showed that it is possible to increase the number of script-specific intrusions by showing subjects several stories that use the same script (e.g. a story about a visit to a doctor and another about a visit to a dentist). Moreover, as Bower et al. put it, "having another version of the same script mention an action increases the probability that the unmentioned analogous action is intruded in recalling a related story." Data from their experiment is shown in Table 3.

The ideas in the sentence processing model can be used with the scripts identified by Bower et al. (1979) to produce a model of their task. Understanding a text means finding the right script to which it corresponds in the background memory. The script plays the role of the context from the sentence processing variant of the model. At study time, when a proposition is read, it is associated with its script analog. At recall, it is plausible that the script will be more active than the propositions that were studied. Moreover, those propositions in the script which corresponded to propositions that were studied should be more active than the others in the script and should have higher probability of retrieval. If several stories based on the same script have been studied, then the script propositions corresponding to the union of propositions in the two texts will be more active than the rest of the script propositions, and therefore intrusions from that set are more likely. Also, because there are more script propositions with higher activation that were not part of the story, there will be a greater number of intrusions. We have implemented this variant of our model in ACT-R and the performance of the model is shown in the last three columns of Table 3.

Conclusions

We have presented a language comprehension model which at each moment attempts to integrate the sentential input with background knowledge. This model offers a unified

explanation for three kinds of empirical phenomena: effects of position on metaphor understanding (Gerrig & Healy, 1983); illusion rates and latencies for good and bad distortions (Ayers et al., 1996) and also latencies and percentage correct in the gist task (Reder & Kusbit, 1991); and finally, text recall of related stories (Bower et al., 1979). In addition to fitting the empirical data, the theory successfully addressed two computational constraints for modeling language comprehension: realistic reaction times — processes subsumed by this theory were accommodated in a time interval on the order of a few seconds, and incremental integration with background knowledge.

References

Anderson, J., & Lebiere, C. (1998). The atomic components of thought. Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Ayers, M., Reder, L., & Anderson, J. (1996). Accepting false information now and believing it later: partial matching and false information in the moses illusion. (Unpublished manuscript)

Bartlett, F. (1932). Remembering: a study in experimental and social psychology. New York & London: Cambridge University Press.

Bower, G., Black, J., & Turner, T. (1979). Scripts in memory for texts. Cognitive Psychology, 11, 177-220.

Bransford, J., & Johnson, M. (1972). Contextual prerequisites for understanding: some inverstigations of comprehension and recall. Journal of Verbal Learning and Verbal Behavior, 11, 717-726.

Erickson, T., & Mattson, M. (1981). From words to meaning: a semantic illusion. Journal of Verbal Learning and Verbal Behavior, 20, 540-552.

Gerrig, R., & Healy, A. (1983). Dual processes in metaphor understanding: comprehension and appreciation. Journal of Experimental Psychology: Memory and Cognition, 9, 667–675.

Oostendorp, H. van, & Kok, I. (1990). Failing to notice errors in sentences. Language and cognitive processes, 5, 105-113.

Oostendorp, H. van, & Mul, S. de. (1990). Moses beats adam: a semantic relatedness effect on a semanticillusion. Acta Psychologica, 74, 35-46.

Ortony, A., Schallert, D., Reynolds, R., & Antos, S. (1978). Interpreting metaphors and idioms: Some effects on comprehension. Journal of Verbal Learning and Verbal Behavior, 17, 465–477.

Owens, J., Bower, G., & Black, J. (1979). The "soap opera" effect in story recall. Memory and Cognition, 7, 185-191.

Reder, L., & Kusbit, G. (1991). Locus of the moses illusion: imperfect encoding, retrieval, or match? Journal of Memory and Language, 30, 385-406.

Sanford, A., & Garrod, S. (1998). The role of scenario mapping in text comprehension. Discourse processes, 26, 159-190.

Searle, J. (1979). Metaphor. In A. Ortony (Ed.), Metaphor and thought. Cambridge University Press.