All Knowledge All the Time:
a Cognitive Science Challenge for Machine Language Comprehension

Kevin Livingston and Christopher K. Riesbeck
EECS Department
Northwestern University
2133 Sheridan Rd. Evanston, IL 60208, USA
{livingston; c-riesbeck}@northwestern.edu

Abstract

Although it is generally accepted that semantic knowledge is important to language comprehension, the question remains as to when this knowledge should be applied in the task of machine language understanding.

This paper discusses two related results in the filed of Cognitive Neuroscience which suggests humans are activating deep semantic and episodic information early in language understanding. This discussion is followed by a discussion of Direct Memory Access Parsing (DMAP), a text reading approach that leverages existing knowledge and performs integration in the early stages of parsing natural language text. DMAP's understanding is driven by memory structures, and it maps immediately and directly to existing knowledge. Machine reading is fundamentally about integrating new knowledge gleaned from reading with existing knowledge, and it seems not only efficient, but cognitively plausible for this process to be started early in the parsing of natural language.

Introduction

Although it is generally accepted that semantic knowledge is important to language comprehension, the question remains as to when this knowledge should be applied in the task of machine language understanding. As engineers interested in producing, large-scale, real-word NLP systems, studying how one such existing system, the human mind, works can provide crucial insights for machine language comprehension. Results in modern Cognitive Neuroscience suggest that in human language understanding, deep semantic and episodic knowledge is used at a very early stage in reading, within hundreds of milliseconds of each word recognition. This presents a challenge to us as AI researchers to produce cognitive models of language understanding that can provide support for using large amount of knowledge quickly and online.

Direct Memory Access Parsing (DMAP) (Martin 1992) is one such response to this challenge. DMAP maps language input directly to existing semantic and episodic knowledge[1], iteratively updating its understanding as it processes text.

Neuroimaging provides techniques which allow cognitive neuroscientists to isolate both when and where in the brain various events are occurring. We will discuss two related studies in the following section, and then discuss DMAP in the section after that.

Neurophysiological Evidence

A series of studies by Camblin et. al. (2007) illustrates the depth of semantic knowledge humans bring to bear early in language understanding. They presented subjects with stories such as the following.

> (I) Lynn had gotten a sunburn at the beach. Nothing she tried would help her dry and irritated skin. Lynn couldn't stop scratching her arms and legs.

> (II) Lynn's wool sweater was uncomfortable and itchy. She fidgeted as the rough material irritated her skin. Lynn couldn't stop scratching her arms and legs.

They found that within 400ms subjects had discerned that "legs" in the context of (II) was anomalous, since sweaters typically shouldn't make ones legs itch.[2] Specifically, Camblin et. al. found that people produced significantly different amplitude N400 ERP potentials when presented with the different stimuli. (Event-related potentials (ERP) are variations in electrical charge, measurable using electroencephalography (EEG), believed to be caused by thoughts. The amplitude of N400 potential is generally accepted to grow as linguistic stimuli become harder to understand.)

This suggests that humans are activating semantic information early in language understanding, and not just shallow understandings, such as that sweaters are itchy, but very detailed information, such as what body parts sweaters can make itchy.

Even more interesting are the results of Hagoort et. al. (2004), who presented subjects with one version of the following sentence.

> (III) The Dutch trains are *(yellow/white/sour)* and very crowded.

---

[1] Here we make the distinction between semantic knowledge i.e. knowledge about ontologies and structure (politicians are humans, humans have two arms, etc.) and episodic knowledge about specific instances (Bill Clinton is a politician, there was a debate last week, etc.).

[2] For comparison having humans read a word out loud can take over 400ms with favorable priming, and closer to 500ms with neutral priming, identifying if a sequence of letters is a word or not a word can take as long as 700-800ms (Schwanenflugel 1991).

There are three versions, where only one of the words "yellow", "white", or "sour" was used. Hagoort reports that it is well known among Dutch people that Dutch trains are yellow. Therefore, the use of *yellow* is understood to be correct. *White*, although a valid modifier for trains, contradicts episodic (world) knowledge. Finally, *sour* is a complete violation of semantic constraints.

"As expected, the classic N400 effect was obtained for the semantic violations. For the world knowledge violations, we also observed a clear N400 effect. Crucially, this effect was identical in onset and peak latency and was very similar in amplitude and topographic distribution to the semantic N400 effect. This finding is strong empirical evidence that lexical semantic knowledge and general world knowledge are both integrated in the same time frame during sentence interpretation, starting at ~300 ms after word onset." (Hagoort 2004)

DMAP

DMAP treats reading text as fundamentally a world knowledge recognition task. DMAP creates new knowledge structures and instances only when it cannot map the input text to existing world knowledge. DMAP uses a collection of language patterns that recursively map textual references to knowledge structures, and even specific instances in memory when possible. Instances include not only individuals like people, George W. Bush, or countries, Iraq, but also events in episodic memory, such as the Madrid train bombings of 2004. Functionally the pattern matcher in DMAP is similar to a tabular chart parser, except our outputs are fully grounded in the underlying knowledge base. Our implementation of DMAP sits on top of ResearchCyc (Lenat 1995), which has over 1.2 million predicate logic assertions.

At every step of the process DMAP accesses the underlying knowledge base. Proper names are immediately turned into references to their related concepts. The KB is queried for bindings (values) for open-ended references, for example, the input text "A bombing occurred in Al Anbar." would immediately send DMAP looking for instances of bombings it already knows about and that have occurred in Al Anbar (Iraq). Followed by the sentence "The attack killed 19 soldiers." DMAP will use the information from the first sentence to help understand the second. Coreference resolution can be performed using the event instances it retrieved from the first sentence, by substituting the bombings it found for the "attack" and attempt to confirm (locate) them in memory.

Methods such as these turn the language comprehension problem into a recognition and confirmation problem. For more discussion see Livingston (2007). This trades some issues in resolving lexical and linguistic ambiguity for a different problem of scale, managing numerous references to the KB. This is an ongoing area of research. The implementations discussed here process text at an average rate of under 2 seconds per sentence. The median time for a sentence is under 0.5 seconds, for our test corpus of 956 sentences, comprising 183 single topic paragraphs.

Summary

The question of when to leverage semantic and episodic knowledge in machine language comprehension systems has been primarily treated in two ways: (1) language understanding is a process of structure building followed by integration with memory, versus (2) language understanding is a process of recognition where parsers access and update memory directly.

This paper discusses results in the field of Cognitive Neuroscience which suggests humans function more like (2). This also provides challenges for machine understanding systems to bring a wealth of real world knowledge to bear early in the understanding process but still operate at scale.

Direct Memory Access Parsing (DMAP) is a text reading approach that leverages existing knowledge and performs integration in the early stages of parsing natural language text. DMAP's understanding is driven by memory structures, and it maps immediately and directly to existing knowledge. Machine reading is fundamentally about integrating new knowledge gleaned from reading with existing knowledge, and it seems not only efficient, but cognitively plausible for this process to be started early in the parsing of natural language.

Acknowledgements

The other members of the Learning Reader research group: Ken Forbus, Larry Birnbaum, Abhishek Sharma, and Leo Ureel. This project was funded through DARPA, grant HR0011-04-1-0051. Thanks to Cycorp for ResearchCyc.

References

Camblin CC, Gordon PC, Swaab TY. 2007 he interplay of discourse congruence and lexical association during sentence processing: Evidence from ERPs and eye tracking. Journal of Memory and Language. 2007 January; 56(1): 103–128.

Hagoort, P., Hald, L., Bastiaansen, M., and Petersson, K. M. 2004 Integration of Word Meaning and World Knowledge in Language Comprehension (16 April 2004) Science 304 (5669), 438.

Lenat, D. B. 1995 *CYC: A Large-Scale Investment in Knowledge Infrastructure.* Communications of the ACM 38 (11): 33–38.

Livingston, K., and Riesbeck, C.K. 2007. Using Episodic Memory in a Memory Based Parser to Assist Machine Reading, Working notes, AAAI Spring Symposium on Machine Reading, AAAI Press.

Martin, C. E. 1992 *Direct Memory Access Parsing.* PhD thesis, Yale University. 1992.

Schwanenflugel, P. J., and White, C. R. (1991). The influence of paragraph information on the processing of upcoming words. Reading Research Quarterly, 26, 160-177.