PROVIDING CONTEXT FOR FREE TEXT INTERPRETATION

Xiaoshan Pan
CIFE, Stanford University
Stanford, CA 94305, USA
xpan@stanford.edu

Hisham Assal, Ph.D.
CADRC, Cal Poly State University
San Luis Obispo, CA 93407, USA
hassal@computer.org

ABSTRACT

The effort to provide general-purpose natural language understanding (NLU) is difficult because of the ambiguity of natural language sentences that can only be cleared if the context is well defined. For the purposes of applications in a specific domain, this problem can be alleviated by offering a model for the application domain as context for interpreting natural language statements. This paper presents architecture and an implementation for this concept and a discussion for future extensions.

Keywords: Context, Ontology, Link Grammar, WordNet, Free text interpretation

1. INTRODUCTION

General-purpose natural language understanding aims at developing software that is capable of taking a natural language statement and interpreting the exact meaning of it. The result of this interpretation may be an action to be taken, new knowledge to be added to the system or a change in the state of the system. Focus on the interpretation result provides direction for the effort in the NLU field.

1.1. Current Efforts in NLU

While a significant quantity of applications which utilize natural language understanding (NLU) have been developed, NLU techniques have yet to reach a mature state of development. Zaenen and Uszkoreit [12] state, “Despite over three decades of research effort, no practical domain-independent parser of unrestricted text has been developed” (p. 110). Ballim et al. [2] claim, “during these last two years no real improvements have been done in achieving full robustness for a NLU system” (p. 2). Some language applications employ annotated phrase-structure grammars (i.e., grammars with hundreds or thousands of rules to describe differing phrase types). Such grammars become difficult to maintain, extend, and reuse as they attain significant size [12]. Statistical parse decision technique is another approach, which is adopted by some language applications, and requires discriminating models of context, which in turn requires annotated training material to adequately estimate model parameters, rendering its development relatively labor-intensive.

1.2. The Problem of Context

The challenge we face stems from the highly ambiguous nature of natural language, that is, meanings of words/phrases in a natural language vary in different contexts. For example, depending on context, the word "plane" could refer to, among other things, an airplane, a geometric object, or a woodworking tool. As another example, the word “can” may be interpreted as either a noun or a verb whose meanings have nothing to do with each other.

Context provides a framework for understanding the meaning of words and statements. As a popularly accepted fact, it is extremely difficult to interpret free text without knowing the corresponding context. This applies to both humans and any NLU computer system.

1.3. Limited Scope for Useful Applications

The attempt to provide general-purpose understanding software requires identifying the proper context for each statement that it aims to process. This task is difficult since context is subjective, which limits the usefulness of such systems. To simplify the task of understanding natural language statements, we provide a given context, where certain concepts are defined and have a specific, unambiguous meaning. The context is provided in the form of an ontology, where concepts and their relationships are defined in explicit terms and specific meaning is assigned for each concept and every relationship.

We then constrain the process of interpreting a given statement to the given context. Terms in the statement are compared to concepts in the ontology, with synonyms considered, to produce a specific understanding.

2. USE OF ONTOLOGY

“In philosophy, ontology is the study of the kinds of things that exist” [3] (p. 20). For an ontology-based computer system, the ontology defines the context of the vocabularies it contains.

2.1. Application Domains Modeling

An ontology serves as a representation vocabulary that provides a set of terms with which to describe the facts in some application domain. Concepts represented by an ontology can usually be clearly depicted by a natural language because the ontology and the natural language function similarly (i.e., describing the world). Most vocabularies used in ontologies are direct subsets of natural languages. For example, a general ontology uses "thing", "entity", and "physical"; a specific ontology uses "tank", "weapon", and "tree". Depending on the construction of the ontology, the meaning of those words in the ontology could remain the same as in natural language, or vary completely.

The meaning of ontological terms that are not derived directly from a natural language can still be captured by a natural language. For example, the word "ALLFRD" is used in the IMMACCS ontology (an ontology developed in the Collaborative Agent Design Research Center, Cal Poly San Luis Obispo, California), and means "friendly force" in English.

2.2. Richness of Relationships in Ontology Represent Context

Context is defined by The American Heritage Dictionary [11] as "the part of a text or statement that surrounds a particular word or passage and determines its meaning".

Minsky [7] argues, "the secret of what anything means to us depends on how we've connected it to all the other things we know". An ontology is not only a collection of vocabularies but also a compendium of rich relationships among the vocabularies, and these relationships establish the meanings for the vocabularies. Some of the typical relationships are, "IS-A" (inheritance), "HAS-A" (composition), "PART-OF" (aggregation). For example, a "Bus" is a type of "Vehicle", a "Bus" has "tires" and an "engine", and a "Bus" is a part of a "transportation system". These relationships among vocabularies describe meanings for each concept and render a context for an application.

In a natural language, a word may have multiple meanings depending on the applicable context. In a computer system, context may be represented and constrained by an ontology. Vocabularies used in an ontology refer only to the context declared by the ontology (i.e., an

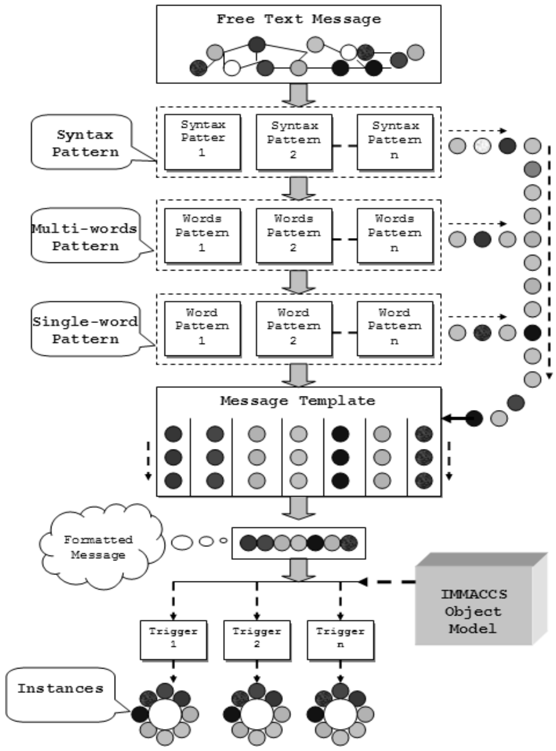

Figure 1: The System Architecture of Early Version of FTI.

ontology provides a context for the vocabulary it contains). When developing an ontology-based computer system, as the actual implementation of associated ontology, a model is constructed to represent the domain knowledge. Thus, software agents are able to disambiguate meanings of words from free text sentences in such a system through utilizing the contextual information provided by the model.

3. FREE TEXT INTERPRETER

The need to provide a 'free text interpreter' in an application came from the need to allow users to interact with the system in a constrained natural language to reduce the need for training and to provide fast response under stress conditions.

3.1. Early Efforts for FTI: the IMMACCS Application

The early version of the Free Text Interpreter (FTI) was developed as part of the extensive capabilities of IMMACCS system, to build an additional communication bridge between human-users and the IMMACCS system (i.e., processing information through free-text messages). FTI established its keyword/key-phrase vocabulary library based on the messages used in real battlefield environments and in the IMMACCS object model, with a high potential of extensibility. It allows a user to input information through the IMMACCS Object Browser (IOB) as free-text messages, and then the system will process these messages with an appropriate interpretation (i.e., creating new objects through the IOB interface with the information derived from the message) and send a feedback message to the user for verification purpose.

Similar to the way in which the human brain processes information in 'chunks', in the FTI a free-text message is viewed as one or more 'chunk(s)' connected in a particular order. A 'chunk' is the smallest unit of meaning, which contains one or more words. It is the task of the FTI to pick up these chunks with their proper associations and instantiate them as IMMACCS objects (Figure 1).

3.2. Lessons Learned from IMMACCS

In this early version of FTI, an initial free text interpretation capability has been successfully achieved while two major constraints were identified:

- FTI has very limited syntactical parsing capability, which precludes it from being a generalized NLU tool.
- Vocabularies that FTI can recognize are only a very small amount compared to the entire English language, which also puts restriction to the usefulness of FTI.

Because each of the above points by itself requires tremendous efforts in terms of research and development, two existing language tools, namely Link Grammar Parser and WordNet database, are included into the new design of a context-based free text interpreter (CFTI).

4. LINK GRAMMAR PARSER

Natural language syntax affects the meaning of words and sentences. The very same words could have different meanings when arranged differently. For example: "a woman, without her man, is nothing" and "a woman: without her, man is nothing" (http://www.p6c.com/Joke_of_the_week.html).

The Link Grammar Parser was found through research to be a very effective syntactic parser, and is therefore incorporated into the design of CFTI.

4.1. Functions of Link Grammar Parser

The Link Grammar Parser, developed at Carnegie Mellon University, is based on "link grammars", an original theory of English syntax [9]. The parser assigns to a given sentence a valid syntactic structure, which consists of a set of labeled links connecting pairs of words.

The Link Grammar Parser utilizes a dictionary of approximately 60,000 word forms, which comprises a significant variety of syntactic constructions, including many considered rare and/or idiomatic. The parser is robust; it can disregard unrecognizable portions of sentences, and assign structures to recognized portions. It is able to intelligently guess, from context and spelling, probable syntactic categories of unknown words. It has knowledge of capitalization, numeric expressions, and a variety of punctuation symbols.

4.2. Basic Concepts of Link Grammar

The basis of the theory of Link Grammar is planarity, described by Melcuk [5], as a phenomenon, evident in most sentences of most natural languages, in which arcs drawn to connect words with specified relationships within sentences do not cross. Planarity is defined in Link Grammar as "the links are drawn above the sentence and do not cross" [9] (p. 7).

Think of words as blocks with connectors coming out. There are different types of connectors; connectors may also point to the right or to the left. A valid sentence is one in which all the words present are used in a way which is valid according to their rules, and which also satisfies certain global rules [10] (p. 1).

Each word, from a Link Grammar perspective, is a block with connectors. Each intricately shaped, labeled box is a connector. A connector is 'satisfied' when 'plugging it into' a compatible connector (as indicated by shape). A valid sentence is one in which all blocks are connected without a cross. The Link Grammar Parser finds out all valid linkages within a free text input, and outputs them as grammatical trees.

5. WORDNET

5.1. The Use of Lexical Database to Relate Concepts

Semantic knowledge concerns what words mean and how these meanings combine in sentences to form sentence meanings [1] (p. 10).

Two types of semantic knowledge are essential in a NLU system: 1) Lexical knowledge among words despite context (e.g., "children" as the plural form of "child", and the synonym relationship between "helicopter" and "whirlybird"); and 2) Contextual

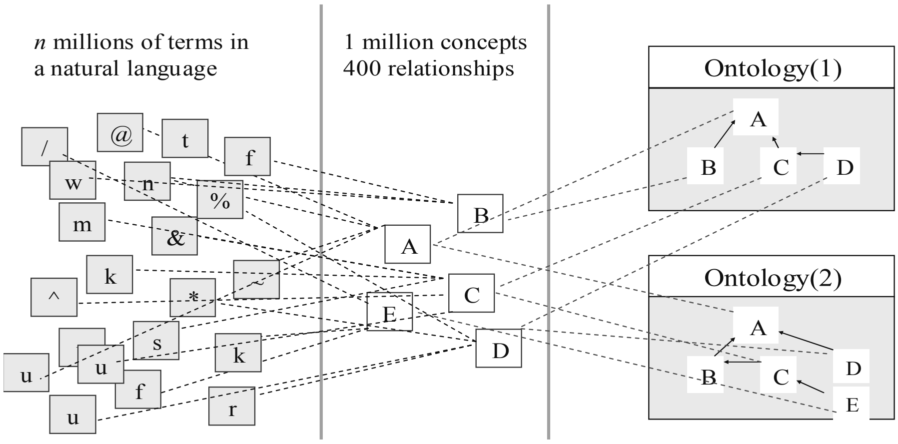

Figure 2: Mapping from natural language to context models.

knowledge (i.e., how meanings are refined when applied to a specified context).

In CFTI, lexical knowledge is acquired through integration of the system with the WordNet database, and contextual knowledge is acquired by tracking contextual meanings of words and phrases during and after development of an ontology (i.e., context model).

WordNet has been developed since 1985 by the Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller. Its design is "…inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, and adjectives are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets." [6] (p. 1).

The most basic semantic relationship in WordNet is synonymy. Sets of synonyms, referred to as synsets, form the basic building blocks. Each synset has a unique identifier (ID), a specific definition, and relationships (e.g., inheritance, composition, entailment, etc.) with other synsets.

WordNet contains a significant quantity of information about the English language. It provides meanings of individual words (as does a traditional dictionary), and also provides relationships among words. The latter is particularly useful in linguistic computing.

While WordNet links words and concepts through a variety of semantic relationships based on similarity and contrast, it "does not give any information about the context in which the word forms and senses occur" [4] (p. 12). In the CFTI, refinement of word meanings in specific contexts (i.e., contextual knowledge) is accomplished by mapping relationships between natural language and a context model.

5.2. Connecting WordNet to an Ontology

Ontologies provide context for vocabularies, which they contain. Direct and indirect mapping relationships exist among ontological vocabularies and natural language vocabularies. Understanding of such relationships may enable a system to understand contextual meanings of words used in the context defined by an ontology. The application of the same word to other ontologies could produce other meanings.

In practice, a process of tracking mapped relationships between a natural language and a context model is a process of interpretation of the model (i.e., what a model really means) through the use of a natural language. The word "ALLFRD", for example, in the IMMACCS ontology means "friendly force" in English. Contextual knowledge can be attained directly from the ontology designer(s), or can be attained through utilization of an automated process if the ontology design follows formalized conventions.

From the perspective of a NLU system, which employs appropriate lexical and contextual knowledge, interpretation of a free text sentence is a process of mapping the sentence from natural language to a context model (Figure 2). Different context models may produce different results simply because words could have different meanings in different contexts.

6. CONTEXT-BASED FTI

The Context-based Free Text Interpreter was designed as a generalized version of the early FTI.

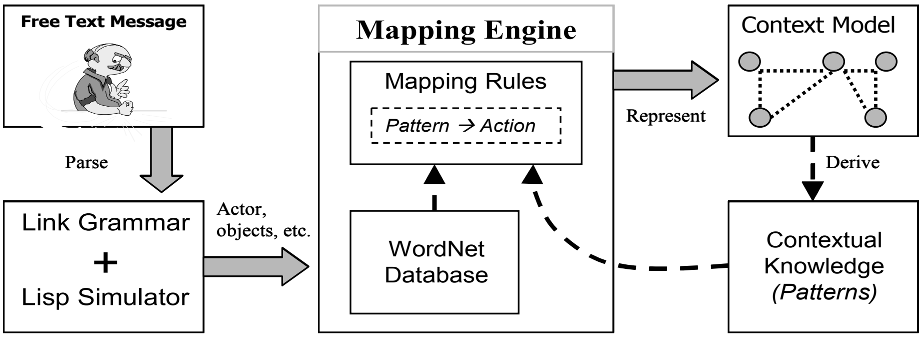

Figure 3: CFTI system architecture.

The motivation behind the generalized design was the need to utilize the FTI tool with different systems within varying context. The main concept in the new design was to separate the concept model from the interpretation mechanism.

6.1. Generalized FTI

In order to interpret a free text sentence correctly, a NLU system needs to conduct the following tasks:

1). Analyze the syntactic structure of the sentence.

2). Analyze the lexical meaning of the words in the sentence.

3). Refine the meanings of the words through the application of a context model.

4). Represent the meaning of the sentence in the model.

Subsequent computations may take place after the above tasks. If an agent engine is attached to the context model, for example, then some agents may react to the changes in the model, if appropriate.

Even though the CFTI system requires an ontological model for the acquisition of contextual knowledge and the representation of meanings, the system is not constrained by any particular knowledge domain. A system change from one ontological model to another does not require significant system reconfigurations.

6.2. Context Building through Ontology Connection

In designing an ontology for a software system, concepts are defined in an unambiguous manner to provide a precise meaning, which is used throughout the system. In addition, relationships are built to connect the concepts and provide a context for the application domain.

6.3. Putting it All Together: System Architecture.

The architecture of the CFTI system consists of five components: Link Grammar, Lisp Simulator, WordNet database, a mapping engine, and a context model (Figure 3).

The Link Grammar is integrated into the CFTI as a syntax parser. With a free text sentence as input, it returns a string, which represents the syntactic structure of the sentence.

Lisp Simulator is a software component implemented in CLIPS [8]. It simulates internal functions (e.g., car, cadr, tail recursion, etc.) that are available in the Lisp language but not available in CLIPS. These functions conveniently facilitate CFTI access to syntactic trees (i.e., outcomes from Link Grammar) and were developed and implemented in CLIPS by the authors.

The WordNet database serves CFTI as a lexical knowledge base, which provides relationships between concepts (i.e., synsets) and words.

A mapping engine is implemented in the form of CLIPS rules. It analyzes the outcome of the Link Grammar and establishes meaning based on the knowledge contained in WordNet (lexical knowledge) and a context model (contextual knowledge).

The context model is a formal ontology created in standard UML (i.e., Unified Modeling Language) format and is subsequently converted into CLIPS COOL classes. A body of domain knowledge is thus represented as a group of classes and relationships. The context model is employed by CFTI for contextual knowledge derivation and meaning representation.

7. CONCLUSION

The general problem of interpreting natural language is difficult due to the ambiguity of words, phrases and sentences. It is essential to have a context, in which the interpretation process is performed. Within a given context, the problem becomes more defined and less ambiguous. In this work we relied on an ontology to provide well-defined context and on existing linguistic tools to help parse the statements and place them in the given context.

7.1. Benefits of Using Ontology

The ontology in this work described the domain knowledge of the given problem (in this case, command and control systems) and defined concepts and their relationships in that domain in unambiguous form for the purposes of this system. Building this ontology provided a context, not only for the system operations, but also for the CFTI tool to perform its function using well-defined concepts.

7.2. The role of WordNet and Link Grammar

The existing linguistic tools, which were used in this work, namely Link Grammar and WordNet, provided valuable assistance in handling the natural language statements. Link Grammar provided powerful mechanism for parsing a statement and presenting it in terms of its components, even if the statement was not a full sentence, i.e. lacking main components, such as a subject. WordNet provided a mechanism to lexically analyze the statement components and place them in collections of synonyms that are related in meaning. These two tools provided the groundwork for analyzing a statement into its main components, maintaining the relationships among the components. This made it possible to build tools to compare the resulting components to the concepts in the ontology using the possible synonyms, if a term in the statement was not directly found in the ontology.

7.3. Future directions and enhancements

Extension of the use of the WordNet database may constitute a potential additional research focus. The WordNet provides the CFTI with lexical knowledge about the English language, however, the current implementation includes only two types of relationships: synonymy and irregular forms of words. Inclusion of additional types of relationships may allow the WordNet to significantly enhance the CFTI's understanding of natural language.

The CFTI system has been tested with a relatively small-sized context model. While an assumption that the system would perform similarly when tested with larger-sized models may be valid, conducting such tests constitutes another potential research focus, which could facilitate the abilities of the CFTI to be used in practical applications.

The CFTI currently assumes that a context model is static, which constrains the system from understanding subjects outside of the model. Future work into extensible ontologies and machine learning would significantly benefit the research study presented in this thesis.

Reference:

[1] Allen, J. (1995). Natural Language Understanding. Redwood City, California: Benjamin/Cummings Publishing Company.

[2] Ballim A., Pallotta V. (1999). Robust Parsing Techniques for Semantic Analysis of Natural Language Queries. VEXTAL'99, November 22-24 1999, Venice - Italy.

[3] Chandrasekaran, B., Josephson, J., Benjamins, V. (1999). What Are Ontologies, and Why do We Need Them? IEEE Intelligent Systems, Vol. 14, No. 1 (January/February 1999), pp. 20-26.

[4] Fellbaum, C. (1999). WordNet: An Electronic Lexical Database. Cambridge: MIT Press.

[5] Melcuk, I. (1988). Dependency Syntax: Theory and Practice. New York: State University of New York Press.

[6] Miller, G. (1990). Wordnet: An Online Lexical Database. Int'l J. Lexicography, Vol. 3, No. 4, 1990, pp. 235-312.

[7] Minsky, M. (1985). The Society of Mind. New York: Simon and Schuster.

[8] NASA. (2002). CLIPS Basic Programming Guide. NASA JSC-25012

[9] Sleator, D., and Temperley, D. (1991). Parsing English with a Link Grammar. Carnegie Mellon University Computer Science technical report CMU-CS-91-196.

[10] Temperley, D., Sleator, D., and Lafferty, J. (1999). [online] An Introduction to the Link Grammar Parser. Technical report, March 1999. Available: http://www.link.cs.cmu.edu/link/

[11] The American Heritage. (2000). Dictionary of the English Language, Fourth Edition. Boston: Houghton Mifflin Company.

[12] Zaenen, A., and Uszkoreit, H. (1996). Language Analysis and Understanding. In Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., and Zue, V., [editors], Survey of the State of the Art in Human Language Technology. Cambridge: Cambridge University Press.