2. Natural Language Processing (NLP)

This section provides a brief history of NLP, introduces some of the main problems involved in extracting meaning from human languages and examines the kind of activities performed by NLP systems.

2.1. Background

Natural language processing systems take strings of words (sentences) as their input and produce structured representations capturing the meaning of those strings as their output. The nature of this output depends heavily on the task at hand. A natural language understanding system serving as an interface to a database might accept questions in English which relate to the kind of data held by the database. In this case the *meaning* of the input (the output of the system) might be expressed in terms of structured SQL queries which can be directly submitted to the database.

The first use of computers to manipulate natural languages was in the 1950s with attempts to automate translation between Russian and English [Locke & Booth]. These systems were spectacularly unsuccessful requiring human Russian-English translators to pre-edit the Russian and post-edit the English. Based on World War II code breaking techniques, they took individual words in isolation and checked their definition in a dictionary. They were of little practical use. Popular tales about these systems cite many mis-translations including the phrase "*hydraulic ram*" translated as "*water goat*".

In the 1960s natural language processing systems started to examine sentence structure but often in an ad hoc manner. These systems were based on pattern matching and few derived representations of meaning. The most well known of these is Eliza [Weisenbaum] though this system was not the most impressive in terms of its ability to extract meaning from language.

Serious developments in natural language processing took place in the early & mid 1970s as systems started to use more general approaches and attempt to formally describe the rules of the language they worked with. LUNAR [Woods 1973] provided an English interface to a database holding details of moon rock samples. SHRDLU [Winograd] interfaced with a virtual robot in a world of blocks, accepting English commands to move the blocks around and answer questions about the state of the world. Since that time there has been parallel development of ideas and technologies that provide the basis for modern natural language processing systems. Research in computer linguistics has provided greater knowledge of grammar construction [Gazdar] and Artificial Intelligence researchers have produced more effective mechanisms for parsing natural languages and for representing meanings [Allen]. Natural language processing systems now build on a solid base of linguistic study and use highly developed semantic representations.

Recently (during the 1990s) natural language systems have either focused on specific, limited domains with some success or attempted to provide general purpose language understanding ability with less success. A major goal in contemporary language processing research is to produce systems which work with complete threads of discourse (with human like abilities) rather than only with isolated sentences [Russell & Norvig(a)]. Successes in this area are currently limited.

2.2. Problems

Two problems in particular make the processing of natural languages difficult and cause different techniques to be used than those associated with the construction of compilers etc for processing artificial languages. These problems are (i) the level of ambiguity that exists in natural languages and (ii) the complexity of semantic information contained in even simple sentences.

Typically language processors deal with large numbers of words, many of which have alternative uses, and large grammars which allow different phrase types to be formed from the same string of words. Language processors are made more complex because of the irregularity of language and the different kinds of ambiguity which can occur. The groups of sentences below are used as examples to illustrate different issues faced by language processors. Each group is briefly discussed in the following section (in keeping with convention, ill-formed sentences are marked with an asterix).

1. The old man the boats.

2. Cats play with string.
   * Cat play with string.

3. I saw the racing pigeons flying to Paris.
   I saw the Eiffel Tower flying to Paris.

4. The boy kicked the ball under the tree.
   The boy kicked the wall under the tree.

1. In the sentence "The old man the boats" problems, such as they are, exist because the word "old" can be legitimately used as a noun (meaning a collection of old people) as well as an adjective, and the word "man" can be used as a verb (meaning take charge of) as well as a noun. This causes ambiguity which must be resolved during syntax analysis. This is done by considering all possible syntactic arrangements for phrases and sub-phrases when necessary.

   The implication here is that any parsing mechanism must be able to explore various syntactic arrangements for phrases and be able to backtrack and rearrange them whenever necessary.