Technical Correspondence

On the Need for Parsing Ill-Formed Input

Kwasny and Sondheimer (1981) present techniques for handling several types of ill-formed input within the context of augmented transition networks. Such approaches are necessary if natural language understanding systems are to handle the full range of input to be expected from users. We present here some statistics which illustrate the magnitude of this problem; almost one third of the queries collected in a natural language processing experiment (McLean, 1981) contain problems of the sort they describe.

They consider three classes of ill-formed input: (1) co-occurrence violations, (2) ellipsis and extraneous terms, and (3) conjunction. A co-occurrence violation results from a problem such as lack of agreement. Ellipsis is the omission of some of the words required for a complete sentence; this problem is grouped together with the problem of extraneous terms, in which unnecessary words are used. The use of conjunction is *not* ungrammatical, but it is included in their classification because similar techniques can be used to handle it.

Kwasny and Sondheimer present limited evidence which indicates the importance of these problems. However, different studies are cited for each of the three problem classes, and not all of them are from the context of natural language understanding systems.

In an experiment designed to test the influence of experience level on the types of queries posed to a natural language understanding system, three groups of student subjects were asked to compose simple English sentences requesting personnel information from a database. Since many of the 693 queries collected were not complete sentences in standard English, the types of errors made were analyzed.

Co-occurrence violations were found in 12.3% of the queries. The most common such problems were incorrectly formed possessives and lack of agreement between subject and verb. Extraneous terms and ellipsis were observed in 14.0% of the queries; the use of ellipsis was far more common than the appearance of extraneous terms. Conjunctions were found in 11.4% of the queries. At least one of these problems was found in 32.8% of the queries. A more detailed breakdown is given in the following table. (It should be noted that these categories are not mutually exclusive.)

A. Co-occurrence violations	85	12.3%
Pronoun/noun disagreement	4	0.6%
Subject/verb disagreement	16	2.3%
Incorrect verb form	2	0.3%
Apostrophe not used in possessive	33	4.8%
Apostrophe used in plural	10	1.4%
Apostrophe misplaced	3	0.4%
Possessive uninflected	12	1.7%
Plural uninflected	5	0.7%
Other co-occurrence violations and grammatical problems	9	1.3%
B. Ellipsis and extraneous terms	97	14.0%
Telegraphic ellipsis	38	5.5%
Missing words		
Articles	23	3.3%
Prepositions	4	0.6%
Other words	8	1.2%
Incomplete sentences	26	3.8%
Extraneous words		
Extra words	2	0.3%
Parenthetical comments	9	1.3%
C. Conjunction	79	11.4%
D. At least one of the above	227	32.8%

C.M. Eastman
Mathematics and Computer Science Dept.
Florida State University
Tallahassee, Florida 32306

D.S. McLean
IBM Corporation
P.O. Box 1328
Boca Raton, Florida 33432

References

Kwasny, Stan C., and Sondheimer, Norman K., "Relaxation Techniques for Parsing Grammatically Ill-Formed Input in Natural Language Understanding Systems", *Am. J. Comp. Ling. 7, 2* (April-June 1981), 99-108.

McLean, D.S., "METASZK: A Natural Language Front End to System 2000," *M.S. Thesis*, Department of Mathematics and Computer Science, Florida State University, Tallahassee, Florida, March 1981.