Blitz: A Preprocessor for Heuristically Detecting Context-Independent Linguistic Structures

Boris Katz, Jimmy Lin & Sue Felshin

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The flow of natural language is often broken by constructions such as numbers, dates, addresses, etc., which are difficult to analyze with conventional linguistic parsers. Blitz, a heuristic-based natural language preprocessor, has been integrated into the START Natural Language System [1], considerably improving START's ability to analyze real-world sentences.

Motivation: Real-world sentences are populated with numerous constructions that do not submit neatly to regular linguistic parsing methods. To handle these constructions, natural language systems typically implement specialized new rules. This leads to a level of complexity which renders development and maintenance difficult. Because these constructions have highly regular forms, and can be largely understood in the absence of context, it is possible to shift the burden of processing away from the primary parser, and onto a simpler, faster, non-linguistic preprocessor.

Previous Work: There already exist several systems (such as [2, 3, 4, 5, 6, 7, 8, 9]) which specialize in the extraction of proper nouns and names. However, the focus of the Blitz system differs somewhat from these other systems. Blitz was designed to handle not only proper nouns, but the entire spectrum of special constructions, and to assist in the goal of natural language understanding, as opposed to previous systems' somewhat less ambitious goals of automatic indexing, keyword extraction, and summary generation. Blitz formerly contained "symbol table" capabilities [10] which have since been transferred to Omnibase [11].

Approach: Blitz uses very simple heuristic rules to extract the above-mentioned constructions from free text and return results in a uniform structure. Ultimately, all information is passed back to START (or any natural language system), endowing it with the ability to understand sentences that it otherwise would not be able to understand.

Blitz employs minimal linguistic and lexical knowledge. It recognizes typographical properties (character position and case) and certain small closed classes of words, e.g., the names of the twelve months, cardinal and ordinal digits, etc., and employs very simple rules for recognizing combined constructions, e.g., a month name and an ordinal represent a date ("June 3rd"). These simple rules do not result in much overgeneration because most special constructions take highly defined forms.

The Blitz system embodies the concept of compartmentalization in its system architecture, isolating each component from another to create independent sections that can easily be interchanged and switched on or off. This architectural design allows Blitz to be specifically adapted to any application, and leads to a system that is easily fine-tuned, maintained, and improved.

Difficulty: Blitz's simple heuristics are extremely effective for recognizing dates, numbers, measures, and other highly restricted expressions. Blitz uses case to recognize proper nouns:

(1) Victor Fleming directed Gone With The Wind.

This is a useful heuristic, but is only partially effective because case is not an absolute indicator for proper nouns. The first word of a sentence is capitalized regardless of whether it is part of a name; note Sentence (2) and Sentence (3). Heuristics can't determine boundaries between proper names; see Sentence (4).

(2) In The New York Times today there was an article about artificial intelligence.
(3) For Better or Worse is a popular comic strip.
(4) The copy of the New York Times John read was missing an entire section.

These problems have been largely solved through the use of symbol tables (see preceding abstract [11]).

Impact: Integrating Blitz with the START Natural Language System has improved START's ability to handle real-world sentences dramatically by allowing it to understand tokens which do not exist in its lexicon. Previously, unknown words and constructions would trigger an interaction such as the one below:

USER: Victor Fleming directed Gone With The Wind. START: Could you phrase that a little differently, I didn't understand.

However, an integrated START system taking advantage of Blitz's pre-processing ability does understand such sentences. In this case, the sentence is passed to the preprocessor, which tokenizes "Victor Fleming" and "Gone With The Wind" as proper nouns. When this information is returned to START, the sentence is transformed into the equivalent of "A directed B," which is then easily parsed.

USER: Victor Fleming directed Gone With The Wind. *START stores this information.* USER: Who directed Gone With The Wind? START: Victor Fleming directed Gone With The Wind.

Future Work: The interaction between START and Blitz needs to be further defined and implemented. A relatively unexplored area is the circumstances in which it would be wise to trust the Blitz interpretation of a sentence more than the interpretation offered by the natural language parser. In certain cases, one can see that Blitz offers a more correct interpretation, but this is non-trivial to determine computationally.

Research Support: This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory.

References:

[1] B. Katz, et al. Information Access Using Natural Language. *MIT Artificial Intelligence Laboratory Research Abstracts, September 2000* (this volume).

[2] NetOwl Extractor Technical Overview. IsoQuest, Inc., March, 1997.

[3] Y. Ravin and N. Wacholder. Extracting Names From Natural-Language Text. IBM Research Report RC 20338, 1997.

[4] D. Bikel, S. Miller, R. Schwartz and R. Weischedel. Nymble: a High-Performace Learning Name-finder. *Proceedings of the Fifth Conference on Applied Natural Language Processing*, 1997.

[5] D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama and M. Tyson. The SRI MUC-5 JV FASTUS Information Extraction System. *Proceedings of the Fifth Message Understanding Conference*, 1993.

[6] P. Hayes. NameFinder: Software that Finds Names in Text. *Proc. of RIAO '94*, Paris, France, 1994.

[7] W. Lehnert, J. McCarthy, S. Soderland, E. Riloff, C. Cardie, J. Peterson and F. Feng. UMASS/HUGHES: Description of the CIRCUS System Used for MUC-5

[8] Y. Ravin, N. Wacholder and M. Choi. Disambiguation of Proper Names in Text. IBM Research Report RC 20735, 1997.

[9] Managing Text with Oracle8 ConText Cartridge Oracle Corporation, Technical White Paper, June, 1997.

[10] B. Katz, et al. Blitz: A Preprocessor for Detecting Context-Independent Linguistic Structures, Part 2, Symbol Tables. *MIT Artificial Intelligence Laboratory Research Abstracts, September 1998.*

[11] B. Katz, et al. Omnibase: A Universal Data Source Interface. *MIT Artificial Intelligence Laboratory Research Abstracts, September 2000* (this volume).