SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	November 30, 1998	Final, 28 September 94 - 31 August 98

4. TITLE AND SUBTITLE
Language Processing Research and Technology - New Directions

5. FUNDING NUMBERS
DAAH04-94-G-0426

6. AUTHOR(S)
Aravind K. Joshi

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
University of Pennsylvania
Department of Computer and Information Science
200 South 33rd Street
Philadelphia, PA 19104-6389

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 33161.27-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A new statistical parser has been developed which works on unconstrained newspaper text with an 89% success rate. It trains in minutes and its success is the highest in the field so far. A new robust parsing technology called "supertagging" was developed based on lexicalized tree-adjoining grammars (LTAG). This parser gives a very fine grain analysis and achieves a 93% success rate for assigning correct supertags to each word. This "almost parsing" becomes actual full parsing in many application domains. New technologies for automated entity based summarization of newspaper articles based on the conference technology developed earlier.

14. SUBJECT TERMS	15. NUMBER IF PAGES	16. PRICE CODE
	6	

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

REPORT DOCUMENTATION PAGE (SF 298)

1. REPORT DATE: November 30, 1998

2. REPORT TYPE AND DATES COVERED: Final, 28 September 94 - 31 August 98

3. TITLE AND SUBTITLE: Language Processing Research and Technology - New Directions

4. FUNDING NUMBERS: DAAH04-94-G-0426

5. AUTHOR(S): Aravind K. Joshi

6. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES):
   University of Pennsylvania
   Department of Computer and Information Science
   200 S. 33rd Street
   Philadelphia, PA 19104-6389

7. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES):
   U.S. Army Research Office
   P.O. Box 12211
   Research Triangle Park, NC 27709-2211

8. SPONSORING/MONITORING AGENCY REPORT NUMBER: 33161-MA

9. ABSTRACT: A new statistical parser has been developed which works on unconstrained newspaper text with an 89% success rate. It trains in minutes and its success is the highest in the field so far. A new robust parsing technology called "supertagging" was developed based on lexicalized tree-adjoining grammars (LTAG). This parser gives a very fine grain analysis and achieves a 93% success rate for assigning correct supertags to each word. This "almost parsing" becomes actual full parsing in many application domains. New technologies for automated entity based summarization of newspaper articles based on the coreference technology developed earlier.

REPORT DOCUMENTATION PAGE (SF 298)
(CONTINUATION SHEET)

10. LIST OF MANUSCRIPTS SUBMITTED OR PUBLISHED UNDER ARO SPONSORSHIP DURING THIS REPORTING PERIOD, INCLUDING JOURNAL REFERENCES:

- Joshi, A. and A. Sakar. Coordination in tree adjoining grammars: formalization and implementation. *Proceedings of the International Conference on Computational Linguistics (COLING)*, Copenhagen, Denmark, August, 1996.
- Steedman, M. "Implications of Binding for Lexicalized Grammars." Revised version of a paper formerly entitled "Binding and Control in CCG and its Relatives," appearing in *Proceedings of the 3rd Workshop on Tree Adjoining Grammars and Related Formalisms,"* Paris, September 1994, Université de Paris 7.
- Joshi, A. *Partial proof, trees, natural deduction and proof-nets*, to appear in *Resource Sensitive Logics*, (ed. C. Reture), Springer-Verlag, 1997.
- Joshi, A. *Centering in Discourse*, (eds. L. Walker, A. K. Joshi, and E. Prince). Oxford University Press, Oxford, to appear in 1997.
- Steedman, M. *Temporality* In J. van Benthem and A. ter Meulen, (eds.), *Handbook of Logic and Language*, Elsevier North Holland, to appear in 1996.
- Steedman, M. *Surface Structure and Interpretation*, Linguistic Inquiry Monograph No.30, MIT Press, 1996.
- Steedman, M. "Representing Discourse Information for Spoken Dialogue Generation." *Proceedings of International Symposium on Spoken Dialogue, International Conference on Spoken Language Processing (held in conjunction with ICSLP-96)*, Philadelphia, Sept 1996, 89-92.
- Di Eugenio, B. and B. Webber. "Pragmatic Overloading in Natural Language Instructions." *Int'l Journal of Expert Systems*, 9(1), 1996.
- Gertner, A. and B. Webber. "A Bias toward Relevance: Recognizing plans where goal minimization fails." *Proceedings 1996 National Conference on Artificial Intelligence (AAAI-96)*, Portland OR, August 1996.
- Cristea D. and B. Webber. "Expectations in an Incremental Approach to Discourse Processing." Submitted to *ACL/EACL '97*, Madrid, Spain.
- Palmer, M. and Joseph Rosenzweig. "Capturing Motion Verb Generalizations with Synchronous TAGs, Proceedings of Association for Machine Translation (AMTA)-96, Montreal, Quebec, October 1-5, 1996.
- Palmer, M. and Joseph Rosenzweig. "Representing Verb Classes in a Lexicalized Tree-Adjoining Grammar, Predicative Forms Workshop associated with ECAI96, Toulouse, France, August 1-2, 1996.
- Han, C., Xia, F. Palmer, M. and J. Rosenzweig. "Capturing Language Specific Constraints on Lexical Selection with Feature-Based Lexicalized Tree-Adjoining Grammars." Proceedings of the International Conference on Chinese Computing, Singapore, June 4-7, 1996.

- Rosenweig, J. "Ask me no more questions, I'll tell you no more lies.." Siglex Workshop: Breadth and Depth of Semantic Lexicons, ACL-96, Santa Cruz, CA, June 28, 1996.
- M. J. Collins. 1996. A New Statistical Parser Based on Bigram Lexical Dependencies. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 184-191.
- Collins, M. "Three Generative, Lexicalised Models for Statistical Parsing". *Proceedings of the 35th Annual Meeting of the ACL (jointly with the 8th Conference of the EACL), Madrid*, 1997.
- Cristea D. and B. Webber. "Expectations in an Incremental Approach to Discourse Processing." *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics* (ACL/EACL '97), Madrid, Spain.
- Hoa Trang Dang, Joseph Rosenzweig, Martha Palmer. "Associating semantic components with Levin classes". In *Proceedings of Interlingua Workshop, MTSUMMIT97*, San Diego, California, Oct 28, 1997.
- Joshi, A. and S. Weinstein. "Formal Systems for Complexity and Control of Inference: a reprise and some hints." Centering in Discourse, (ed. A. K. Joshi, E. Prince, L. Walker), Oxford University Press, Oxford, 1997.
- Joshi, A. *Centering in Discourse*, (eds. L. Walker, A. K. Joshi, and E. Prince). Oxford University Press, Oxford, 1997.
- Alexis Nasr, Owen Rambow, Martha Palmer, Joseph Rosenzweig. "Enriching lexical transfer with cross-linguistic semantic features". In *Proceedings of the Interlingua Workshop at the MT Summit*, San Diego, California, Oct 28, 1997.
- Stone, M. and C. Doran. "Sentence Planning as Description Using Tree Adjoining Grammar", ACL-EACL97, Madrid, Spain, July 1997.

11. SCIENTIFIC PERSONNEL SUPPORTED BY THIS PROJECT AND DEGREES AWARDED DURING THIS REPORTING PERIOD:

Aravind Joshi, Mitch Marcus, Mark Steedman, Bonnie Webber (Faculty)
Martha Palmer (Senior Research Associate)
Srinivas Bangalore, Hoa Trang Dang, Christy Doran, Jason Eisner, Nobo Komagato, Seth Kulick, Jeff Reynar, Anoop Sakar, William Schuler, Matthew Stone, Fei Xia (students)

Degrees Awarded:

M.S. Degrees awarded: 5
Ph.D. Degrees awarded: 6

Honors/Awards Received:

- Aravind K. Joshi – Recipient of the Research Excellence Award for 1997 by the IJCAI, International Joint Conference of Artificial Intelligence to be held in Nagoya, Japan.
- Elected Fellow of ACM (Association for Computing Machinery), 1997.

- Mitch Marcus – Appointed President, Association for Computational Linguistics, 1997.
- Bonnie Webber – Appointed Program Chair for the 1997 meeting of the American Association for Artificial Intelligence, (AAAI).
- REPORT OF INVENTIONS: none
- SCIENTIFIC PROGRESS AND ACCOMPLISHMENTS:

  The overall research objectives and motivations for our research are concerned with the following general problems: (1) Seamless integration of lexical information with syntactic, semantic, and prosodic (intonational) information. This integration is to be achieved both on a large scale as well as with respect to the complexity of phenomena considered; (2) Systematic and computationally tractable ways of integrating structural and statistical information using large amounts of data, leading to automatic acquisition of various aspects of language structure; and (3) Integration of natural language processing work in discourse with planning, especially in the environment of animated agents.

  Our research continues to be guided by the need to integrate different ways of approaching language processing - language structure, language understanding and generation, acquiring language, and language in real world situations, among others. Experimentation with very large annotated databases or raw text and speech databases is an important aspect of our approach.

Significant Accomplishments.

- Developed a new statistical parser which outperforms any existing parser on unconstrained newspaper text with more than 89% success rate. Unlike earlier statistical parsers, it trains in minutes, rather than months, and parses at 200 sentences per minute. Such parsers hold the promise of major breakthroughs in language modeling for speech recognition.
- Developed a novel robust parsing strategy, called "supertagging", based on treating tree fragments as "super" parts of speech, using a trigram supertagging model with enhanced smoothing techniques and achieving a very high performance by assigning correct supertags to each word with an accuracy exceeding 93%, and about 85% with parsing with respect to Penn Treebank.
- Developed a new set of high performance NLP components (all now the best in the field at this time) for use in message understanding. Applied maximum entropy techniques to yield a new part of speech tagger (see Technology Transition), and an end of sentence detector. Applied Penn's Error Based Transformation Learner (see Tech Transition) to develop world's most accurate Noun Phrase chunker.
- Developed a new system for automatic summarization of newspaper articles, based on a new system to determine which entities co-refer in newspaper text (i.e. that "He" in one sentence refers to the same entity as "Mr. Gates" in another sentence). The first version of the coreference system was entered in the MUC-6 competition; with one bug fix, it outperforms any other system in that competition.
- Incorporated IBIS spoken response system into the TRAUMAID medical expert system, which required the development of information-structured end discourse-planning components.

- Meaning to speech capabilities were developed further in respect to intonation and facial animation using a small combinatory categorical grammar.
- Continued the development of statistical dependency parser with superior performance on unconstrained newspaper text. This new parser uses more linguistic information which is just adequate for estimating the statistics for dependencies. This parser performs at 89% accuracy.
- Substantial further development of the robust parsing strategy, called "supertagging", based on treating tree fragments as "super" part of speech, using a trigram supertagging model with enhanced smoothing techniques and achieving a very high performance by assigning correct supertags to each word with an accuracy exceeding 93%, and about 85% with parsing with respect to Penn Treebank. This parser has been in applications such as information extraction and machine translation.
- Continued the development of a system for automatic summarization of newspaper articles, based on a new system to determine which entities co-refer in newspaper text (e.g., "He" in one sentence may refer to the same entity as "Mr. Gates" in another sentence). This system will be entered in the MUC-7 competition in October 98.
- Developed new techniques for creating classes of lexical items based on a novel technique called 'intersective classes'. This verb classification has application to machine translation.

• TECHNOLOGY TRANSFER:

- Lexis-Nexus: supported research and development in information extraction and summarization.
- Lockheed Martin: Coreference and information extraction.
- Digital Equipment Corporation: summarization and translation of service reports.
- Sun Micro Systems: Extraction of linguistic resources from aligned bilingual corpa. Donation of high performance computers for statistical NLP.
- IBM, BBN and AT&T Laboratories: Provide summer opportunities for advanced graduate students.
- Cogentex, Ithaca: SBR, Phase I Award. Now developing Phase II proposal.