Current Status & Process in the Development of Applications Through NLP

V R Rathod S M Shah Nileshkumar K Modi

Abstract

The development of natural language processing systems has resulted in their being increasingly used in support of other computer programs. This trend is particularly noticeable with regard to information management applications. Natural language processing provides a potential means of gaining access to the information inherent in the large amount of text and available through the Internet. In the following survey, we look in further details at the recent trends in research in natural language processing and conclude with a discussion of some applications of this research to the solution of information management problems.

Keywords : Natural Language Processing.

0. Introduction

Work in computational linguistics began very soon after the development of the first computers, yet in the intervening four decades there has been a pervasive feeling that progress in computer understanding of natural language has not been commensurate with progress in other computer applications. Recently, a number of prominent researchers in natural language processing met to assess the state of the discipline and discuss future directions. The consensus of this meeting was that increased attention to large amounts of lexical and domain knowledge was essential for significant progress, and current research efforts in the field reflect this point of view.

1. Passive Voice and Its Usage

The traditional approach in computational linguistics included a prominent concentration on the formal mechanisms available for processing language, especially as these applied to syntactic processing and, somewhat less so, to semantic interpretation. In recent efforts, work in these areas continues, but there has been a marked trend toward enhancing these core resources with statistical knowledge acquisition techniques. There is considerable research aimed at using online resources for assembling large knowledge bases, drawing on both natural language corpora and dictionaries and other structured resources. Recent research in lexical semantics reflects an interest in the proper structuring of this information to support linguistic processing.

Furthermore, the availability of large amounts of machine-readable text naturally supports continued work in analysis of connected discourse. In other trends the use of statistical technique are being used as part of the parsing process, for automatic part of speech assignment, and for word sense disambiguation.

2. The Lexicon

In computational linguistics the lexicon supplies paradigmatic information about words, including part of speech labels, irregular plurals, and sub categorization information for verbs. Traditionally, lexicons were quite small and were constructed largely by hand. There is a growing realization that effective natural language processing requires increased amounts of lexical (especially semantic) information. A recent trend has been the use of automatic techniques applied to large corpora for the purpose of acquiring lexical information from text. Statistical techniques are an important aspect of automatically mining lexical

information. Manning (1993) uses such techniques to gather sub categorization information for verbs. Brent (1993) also discovers Sub categorization information; in addition he attempts to automatically discover verbs in the text. Liu and Soo (1993) describe a method for mining information about thematic roles.

The additional information being added to the lexicon increases the complexity of the lexicon. This added complexity requires that attention be paid to the organization of the lexicon: Zernik 1991 (Part III) and Pustejovsky 1993 (Part III) both contain several papers which address this issue. McCray, Srinivasan and Browne(1993) discuss the structure of a large (more than 60,000 base forms) lexicon designed and implemented to support syntactic processing.

3. Automatic Tagging

Automatically disambiguating part-of-speech labels in text is an important research area since such ambiguity is particularly prevalent in English. Programs resolving part-of-speech labels (often called automatic taggers) typically are around 95% accurate. Taggers can serve as preprocessors for syntactic parsers and contribute significantly to efficiency. There have been two main approaches to automatic tagging: probabilistic and rule-based. Merialdo (1994) and Dermatos and Kokkinakis (1995) review several approaches to probabilistic tagging and then offer new proposals. Typically, probabilistic taggers are trained on disambiguated text and vary as to how much training text is needed and how much human effort is required in the training process. (See 3 Schütze 1993 for a tagger that requires very little human intervention.) Further variation concerns knowing what to do about unknown words and the ability to deal with large numbers of tags.

One drawback to stochastic taggers is that they are very large programs requiring considerable computational resources. Brill (1992) describes a rule-based tagger which is as accurate as stochastic taggers, but with a much smaller program. The program is slower than stochastic taggers, however. Building on Brill's approach, Roche and Schabes (1995) propose a rule-based, finite-state tagger which is much smaller and faster than stochastic implementations. Accuracy and other characteristics remain comparable.

4. Parsing

The traditional approach to natural language processing takes as its basic assumption that a system must assign a complete constituent analysis to every sentence it encounters. The methods used to attempt this are drawn from mathematics, with context-free grammars playing a large role in assigning syntactic constituent structure. Partee, ter Meulen and Wall (1993) provide an accessible introduction to the theoretical constructs underlying this approach, including set theory, logic, formal language theory, and automata theory, along with the application of these mechanisms to the syntax and semantics of natural language.

The program described in Alshawi 1992 is a very good example of a complete system-built on these principles. For syntax, it uses a unification-based implementation of a generalized phrase structure grammar (Gazdar *et al.* 1985) and handles an impressive number of syntactic structures which might be expected to appear in "interactive dialogues with information systems… although of course there is still a large residue even of this variety of English that the system fails to analyze properly." (Alshawi 1992:61).

In continuing research in this tradition, context-free grammars have been extended in various ways. The so-called "mildly context sensitive grammars," such as tree adjoining grammars, have had considerable influence on recent work concerned with the formal aspects of parsing natural language.

Several recent papers pursue nontraditional approaches to syntactic analysis. One such technique is partial, or underspecified, analysis. For many applications such an analysis is entirely sufficient and can often be more reliably produced than a fully specified structure. Chen and Chen (1994), for example, employ statistical methods combined with a finite state mechanism to impose an analysis which consists only of noun phrase boundaries, without specifying their complete internal structure or their exact place in a complete tree structure. Agarwal and Boggess (1992) successfully rely on semantic features in a partially specified syntactic representation for the identification of coordinate structures. In an innovative application of dependency grammar and dynamic programming techniques, Kurohashi and Nagao (1994) address the problem of analyzing very complicated coordinate structures in Japanese.

A recent innovation in syntactic processing has been investigation into the use of statistical techniques. (See Charniak 1993 for an overview of this and other statistical applications.) In probabilistic parsing, probabilities are extracted from a parsed corpus for the purpose of choosing the most likely rule when more than one rule can apply during the course of a parse (Magerman and Weir 1992). In another application of probabilistic parsing the goal is to choose the (semantically) best analysis from a number of syntactically correct analyses for a given input (Briscoe and Carroll 1993, Black, Garside and Leech 1993).

A more ambitious application of statistical methodologies to the parsing process is grammar induction where the rules themselves are automatically inferred from a bracketed text; however, results in the general case are still preliminary. Pereira and Schabes (1992) discuss inferring a grammar from bracketed text relying heavily on statistical techniques, while Brill (1993) uses only modest statistics in his rule-based method.

5. Word-Sense Disambiguation

Automatic word-sense disambiguation depends on the linguistic context encountered during processing. McRoy (1992) appeals to a variety of cues while parsing, including morphology, collocations, semantic context, and discourse. Her approach is not based on statistical methods, but rather is symbolic and knowledge intensive. Statistical methods exploit the distributional characteristics of words in large texts and require training, which can come from several sources, including human intervention. Gale, Church and Yarowsky (1992) give an overview of several statistical techniques they have used for word-sense disambiguation and discuss research on evaluating results for their systems and others. They have used two training techniques, one based on a bilingual corpus, and another on *Roget's Thesaurus.* Justeson and Katz (1995) use both rule based and statistical methods. The attractiveness of their method is that the rules they use provide linguistic motivation.

6. Semantics

Formal semantics is rooted in the philosophy of language and has as its goal a complete and rigorous description of the meaning of sentences in natural language. It concentrates on the structural aspects of meaning. Chierchia and McConnell-Ginet (1990) provide a good introduction to formal semantics. The papers in Rosner and Johnson 1992 discuss various aspects of the use of formal semantics in computational linguistics and focus on Montague grammar (Montague 1974), although Wilks (1992) dissents from the prevailing view. King (1992) provides an overview of the relation between formal semantics and computational linguistics. Several papers in Rosner and Johnson discuss research in the situation semantics paradigm (Barwise and Perry 1983), which has recently had wide influence in computational linguistics, especially in discourse processing. See Alshawi 1992 for a good example of an implemented (and eclectic) approach to semantic interpretation.

Lexical semantics (Cruse 1986) has recently become increasingly important in natural language processing. This approach to semantics is concerned with psychological facts associated with the meaning of words. Levin (1993) analyzes verb classes within this framework, while the papers in Levin and Pinker 1991 explore additional phenomena, including the semantics of events and verb argument structure. A very interesting application of lexical semantics is WordNet 5 (Miller 1990), which is a lexical database that attempts to model cognitive processes. The articles in Saint-Dizier and Viegas 1995 discuss psychological and foundational issues in lexical semantics as well as a number of aspects of using lexical semantics in computational linguistics.

Another approach to language analysis based on psychological considerations is cognitive grammar (Langacker 1988). Olivier and Tsujii (1994) deal with spatial prepositions in this framework, while Davenport and Heinze (1995) discuss more general aspects of semantic processing based on cognitive grammar.

7. Discourse Analysis

Discourse analysis is concerned with coherent processing of text segments larger than the sentence and assumes that this requires something more than just the interpretation of the individual sentences. Grosz, Joshi and Weinstein (1995) provide a broad-based discussion of the nature of discourse, clarifying what is involved beyond the sentence level, and how the syntax and semantics of the sentences support the structure of the discourse. In their analysis, discourse contains linguistic structure (syntax, semantics), attentional structure (focus of attention), and intentional structure (plan of participants) and is structured into coherent segments. During discourse processing one important task for the hearer is to identify the referents of noun phrases. Inferencing is required for this identification. A coherent discourse lessens the amount of inferencing required of the hearer for comprehension. Throughout a discourse the particular way that the speaker maintains “focus of attention” or “centering” through choice of linguistic structures for referring expressions is particularly relevant to discourse coherence.

Other work in computational approaches to discourse analysis has focused on particular aspects of processing coherent text. Hajicova, Skoumalova and Sgall (1995) distinguish topic (old information) from focus (new information) within a sentence. Information of this sort is relevant to tracking focus of attention. Lappin and Leass (1994) are primarily concerned with intrasentential anaphora resolution, which relies on syntactic, rather than discourse, cues. However, they also address intersentential anaphora, and this relies on several discourse cues, such as saliency of an NP, which is straightforwardly determined by such things as grammatical role, frequency of mention, proximity, and sentence recency. Huls, Bos and Claasen (1995) use a similar notion of saliency for anaphora resolution and resolve deictic expressions with the same principles. Passonneau and Litman (1993) study the nature of discourse segments and the linguistic structures which cue them. Sonderland and Lehnert (1994) investigate machine learning techniques for discovering discourse-level semantic structure.

Several recent papers investigate those aspects of discourse processing having to do with the psychological state of the participants in a discourse, including, goals, intentions, and beliefs: Asher and Lascarides (1994) investigate a formal model for representing the intentions of the participants in a discourse and the interaction of such intentions with discourse structure and semantic content. Traum and Allen (1994) appeal to the notion of social obligation to shed light on the behavior of discourse. Wiebe (1994) investigates psychological point of view in third person narrative and provides an insightful algorithm for tracking this phenomenon in text. The point of view of each sentence is either that of the narrator or any one of the characters in the narrative.6 Wiebe discusses the importance of determining point of view for a complete understanding of a text, and discusses how this interacts with other aspects of discourse structure.

8. Applications

As natural language processing technology matures, it is increasingly being used to support other computer applications. Such use naturally falls into two areas, one in which linguistic analysis merely serves as an interface to the primary program, and another in which natural language considerations are central to the application.

Natural language interfaces to data base management systems (e.g. Bates 1989) translate users' input into a request in a formal data base query language, and the program then proceeds as it would without the use of natural language processing techniques. It is normally the case that the domain is constrained and the language of the input consists of comparatively short sentences with a constrained set of syntactic structures.

The design of question answering systems is similar to that for interfaces to data base management systems. One difference, however, is that the knowledge base supporting the question answering system does not have the structure of a data base. See, for example Kupiec 1993, where the underlying knowledge base is an on-line encyclopedia. Processing in this system not only requires a linguistic description for users' requests, but it is also necessary to provide a representation for the encyclopedia itself. As with the interface to a DBMS, the requests are likely to be short and have a constrained syntactic structure. Lauer, Peacock and Graesser (1992) provide some general considerations concerning question answering systems and describe several applications.

In message understanding systems, a fairly complete linguistic analysis may be required, but the messages are relatively short and the domain is often limited. Davenport and Heinze (1995) describe such a system in a military domain. See Chinchor, Hirschman and Lewis 1993 for an overview of some recent message understanding systems.

In three closely related applications (information filtering, text categorization, and automatic abstracting), no constraints on the linguistic structure of the documents being processed can be assumed. One mitigating factor, however, is that effective processing may not require a complete analysis. For all of these applications there are also statistically based systems based on frequency distributions of words. These systems work fairly well, but most people feel that for further improvements, and for extensions, some sort of understanding of the texts, such as that provided by linguistic analysis, is required.

Information filtering and text categorization are concerned with comparing one document to another. In both applications, natural language processing imposes a linguistic representation on each document being considered. In text categorization a collection of documents is inspected and all documents are grouped into several categories based on the characteristics of the linguistic representations of the documents. Blosseville *et al.* (1992) describe an interesting system which combines natural language processing, statistics, and an expert system. In information filtering, 7 documents satisfying some criterion are singled out from a collection. Jacobs and Rau (1990) discuss a program which imposes a quite sophisticated semantic representation for this purpose.

In automatic abstracting, a summary of each document is sought, rather than a classification of a collection. The underlying technology is similar to that used for information filtering and text categorization: the use of some sort of linguistic representation of the documents. Of the two major approaches, one (e.g. McKeown and Radev 1995) puts more emphasis on semantic analysis for this representation and the other (e.g. Paice and Jones 1993), less.

Information retrieval systems typically allow a user to retrieve documents from a large bibliographic database. During the information retrieval process a user expresses an information need through a query. The system then attempts to match this query to those documents in the database which satisfy the user's information need. In systems which use natural language processing, both query and documents are transformed into some sort of a linguistic structure, and this forms the basis of the matching. Several recent information retrieval systems employ varying levels of linguistic representation for this purpose. Sembok and van Rijsbergen (1990) base their experimental system on formal semantic structures, while Myaeng, Khoo and Li (1994) construct lexical semantic structures for document representations. Strzalkowski (1994) combines syntactic processing and statistical techniques to enhance the accuracy of representation of the documents. In an innovative approach to document representation for information retrieval, Liddy *et al (*1995) use several levels of linguistic structure, including lexical, syntactic, semantic, and discourse.

9. References

1. Allen, J. 1987. Natural language understanding. Menlo Park, CA: The Benjamin/Cummings Publishing Company, Inc.
2. Bates, M. and R. M. Weischedel (eds.) 1993. Challenges in natural language processing. Cambridge: Cambridge University Press. 8
3. Rosner, M. and R. Johnson (eds.) 1992. Computational linguistics and formal semantics. Cambridge: Cambridge University Press.
4. Saint-Dizier, P. and E. Viegas (eds.) 1995. Computational lexical semantics. Cambridge: Cambridge University Press.
5. Wiebe, J. M. 1994. Tracking point of view in narrative. Computational Linguistics 20.2.233-287.
6. Agarwal, R. and L. Boggess. 1992. A simple but useful approach to conjunct identification. In Proceedings of the 30th annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 15-21.
7. Alshawi, H. (ed.) 1992. The core language engine. Cambridge, MA: The MIT Press. Asher, N. and A. Lascarides. 1994. Intentions and information in discourse. In Proceedings of the
8. 32nd annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 34-41.
9. Barwise, J. and J. Perry. 1983. Situations and attitudes. Cambridge, MA: The MIT Press.
10. Bates, M. 1989. Rapid porting of the Parlance Natural Language Interface. In Proceedings of the speech and natural language workshop. San Mateo, CA: Morgan Kaufmann Publishers. 83-88.
11. Black, E., R. Garside and G. Leech (eds.) 1993. Statistically-driven computer grammars of English: The IBM/Lancaster approach. Amsterdam: Editions
12. Rodopi. Blosseville, M.J., et al. 1992. Automatic document classification: Natural language processing, statistical analysis, and expert system techniques used together.
13. N. Belkin, P. Ingwesen and A. M. Pejtersen (eds.) Proceedings of the 15th annual international ACM SIGIR conferenceon research and development in information retrieval. NewYork: Association for Computing Machinery. 51-58.
14. Booth, A. D., L Brandwood and J. P. Cleave. 1958. Mechanical resolution of linguistic problems. London: Butterworths Scientific Publications.

15. Brent, M. R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. Computational Linguistics 19.2.243-262.

16. Brill, E. 1992. A simple rule-based part of speech tagger. In Proceedings of the third conference on applied natural language processing.

17. Trento, Italy. San Francisco: Morgan Kaufmann Publishers. 152-155. 1993. Automatic grammar induction and parsing free text: A transformation-based approach. In Proceedings of the 31st annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 259-265.

18. Briscoe, T. and J. Carroll. 1993. Generalized probabilistic LR parsing of natural language (corpora) with unification-based grammars. Computational Linguistics 19.1.25-59.

19. Charniak, E. 1993. Statistical language learning. Cambridge, MA: The MIT Press.

20. Chierchia, G. and S. McConnell-Ginet. 1990. Meaning and grammar: An introduction to semantics. Cambridge, MA: The MIT Press.

21. Chinchor, N., L. Hirschman and D. D. Lewis. 1993. Evaluating message understanding systems: An analysis of the Third Message Understanding Conference (MUC-3). Computational Linguistics 19.3.409-450.

22. Cruse, D. A. 1986. Lexical semantics. Cambridge: Cambridge University Press.

23. Davenport, D. M. and D. T. Heinze. 1995. Crisis action message analyzer - EDM. Proceedings of the 5th annual dual-use technologies and applications conference. SUNY Institute of Technology at Utica/Rome, NY. 284-289.

24. Dermatas, E. and G. Kokkinakis. 1995. Automatic stochastic tagging of natural language texts. Computational Linguistics 21.2.137-163.

25. Fries, U., G. Tottie and P. Schneider (eds.) 1994. Creating and using English language corpora: Papers from the fourteenth international conference on English language research on computerized corpora, Zurich 1993. Amsterdam: Editions Rodopi.

26. Gale, W., K. W. Church and D. Yarowsky. 1992. Estimating upper and lower bounds on performance of word-sense disambiguation programs. In Proceedings of the 30th annual meeting of the Association for Computational Linguistics. San Francisco: Morgan Kaufmann Publishers. 249-256.

27. Gazdar, G., et al. 1985. Generalized phrase structure grammar. Oxford: Blackwell Publishing and Cambridge, MA: Harvard University Press.

28. Grosz, B. J., A. K. Joshi and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 21.2.203-225.

29. Hajicova, E., H. Skoumalova and P. Sgall. 1995. An automatic procedure for topic-focus identification. Computational Linguistics 21.1.81-94.

30. Huls, C., E. Bos and W. Claasen. 1995. Automatic referent resolution of deictic and anaphoric expressions. Computational Linguistics 21.1.59-79.

31. Jacobs, P. S. and L. F. Rau. 1990. SCISOR: Extracting information from on-line news. Communications of the ACM 33.11.88-97.

32. Justeson, J. S. and S. M. Katz. 1995. Principled disambiguation: Discriminating adjective senses with modified nouns. Computational Linguistics 21.1.1-27.

33. King, M. 1992. Epilogue: On the relation between computational linguistics and formal semantics.

34. In M. Rosner and R. Johnson (eds.) Computational linguistics and formal semantics. Cambridge: Cambridge University Press. 283-299.

35. Kupiec, J. 1993. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia.

36. In R. Korfhage, E. Rasmussen and P. Willett (eds.) Proceedings of the 16th annual international ACM SIGIR conference on research and development in informationretrieval. New York: Association for Computing Machinery. 181-190.

37. Kurohashi, S. and M. Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. Computational Linguistics 20.4.507-534.

38. Langacker, R. W. 1988. An overview of cognitive grammar. In B. Rudzka-Ostyn (ed.) Topics in cognitive linguistics. Amsterdam/Philadelphia: John Benjamins Publishing Company. 3- 48.

About Authors

Dr. V R Rathod is a Professor & Head in Department of Computer Science, Bhavnagar University, Bhavnagar, Gujarat.
E-mail : profvrr@rediffmail.com

Prof. S M Shah is a Director in S. V. Institute of Computer Studies, S. V. Campus, Kadi, Gujarat.
E-mail : prof_smshah@yahoo.com

Mr. Nileshkumar K Modi is a Lecturer in S. V. Institute of Computer Studies, S. V. Campus, Kadi, Gujarat.
E-mail : tonileshmodi@yahoo.com