Natural language processing: a historical review

Karen Sparck Jones
Computer Laboratory, University of Cambridge
William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, UK
sparckjones@cl.cam.ac.uk, http://www.cl.cam.ac.uk/~ksj

October 2001

Abstract

This paper reviews natural language processing (NLP) from the late nineteen forties to the present, seeking to identify its successive trends as these reflect concerns with different problems or the pursuit of different approaches to solving these problems and building systems as wholes. The review distinguishes four phases in the history of NLP, characterised respectively by an emphasis on machine translation, by the influence of artificial intelligence, by the adoption of a logico-grammatical style, and the use of massive language data. The account considers the significant and salient work in each phase, and concludes with an assessment of where we stand after more than fifty years of effort in the field. [1]

1 Introduction

Work in the NLP field has concentrated first on one problem, then on another, sometimes because solving problem X depends on solving problem Y but sometimes just because problem Y seems more tractable than problem X, or because there is market interest in a solution to Y. There has been very substantial progress, both in understanding how to do NLP and in actually doing it, since work in the field took off in the 1950s. In the last fifteen years in particular, advances in computing technology have made it possible to implement ideas that could only be adumbrated before, to consolidate research, and to carry speech and language processing into the ordinary world. Sometimes the scientific advance in NLP, or the computational linguistics underlying it, is less than the onward rush of information technology is taken to imply; but the theoretical and empirical development needed to establish a new field is evident in the fifty-year period reviewed here. Sometimes innovation is only old ideas reappearing in new guises, like lexicalist approaches to NLP, or shallow parsing. But the new costumes are better made, of better materials, as well as more becoming: so the research is not so much going round in circles as ascending a spiral.

In this brief review, I divide the history of NLP into four phases, with distinctive concerns and styles. I define the first phase of work in NLP as lasting from the late 1940s to the

[1] This paper uses material from my article 'Natural language processing: an overview', in the first and second editions of *International encyclopedia of linguistics*, New York: Oxford University Press, 1992, (Ed. W. Bright), and in press (Ed. W. Frawley), and from 'Natural language processing: a historical overview' in *Current issues in computational linguistics: in honour of Don Walker*, Amsterdam: Kluwer, 1994, (Ed. A. Zampolli, N. Calzolari and M. Palmer).

late 1960s, the second from the late 60s to the late 70s and the third to the late 80s, with the fourth phase to the end of the century. The first phase was driven by MT, the second flavoured by AI, the third grammatico-logical, while the fourth phase has focused on lexical and corpus data. This last phase illustrates the cyclic character of work in the field, since it has returned to a major concern of the first period, namely building the powerful and comprehensive dictionaries that serious NLP applications like MT need, and has revived early proposals for extracting linguistic information from large natural corpora. At the same time, it also illustrates the extent to which real progress has been made in the field in such key areas as computational grammar and parsing.

2 Phase 1: late 1940s to late 1960s

The work of the first phase was focused on machine translation (MT). Following a few early birds, including Booth and Richens' investigations and Weaver's influential memorandum on translation of 1949 (Locke and Booth, 1955), research on NLP began in earnest in the 1950s. Automatic translation from Russian to English, in a very rudimentary form and limited experiment, was exhibited in the IBM-Georgetown Demonstration of 1954. The journal *MT (Mechanical Translation)*, the ancestor of *Computational Linguistics*, also began publication in 1954. The first international conference on MT was held in 1952, the second in 1956 (the year of the first artificial intelligence conference); at the important Washington International Conference on Scientific Information of 1958 language processing was linked with information retrieval, for example in the use of a thesaurus, Minsky drew attention to artificial intelligence, and Luhn provided auto-abstracts (actually extracts) for one session's papers. The Teddington International Conference on Machine Translation of Languages and Applied Language Analysis in 1961 was perhaps the high point of this first phase: it reported work done in many countries on many aspects of NLP including morphology, syntax and semantics, in interpretation and generation, and ranging from formal theory to hardware.

This first phase was a period of enthusiasm and optimism. It is notable not only because those engaged attacked a very difficult NLP task, and so encountered the problems of syntactic and semantic processing, and of linguistic variety, in all their force; they were seeking to use a new tool, computers, for non-numerical, data-processing purposes when data-processing itself was not well established. It is essential to remember how primitive the available computing resources were. This was the era of punched cards and batch processing. There were no suitable higher-level languages and programming was virtually all in assembler. Access to machines was often restricted; they had very limited storage, and were extremely slow. Plath (1967) reports processing speeds like 7 minutes for analysing long sentences, even with the most advanced algorithms and on the best machines then available. Vast amounts of programming effort were devoted to bit-packing to save space and time. It is remarkable how much was done with such poor resources, for example in grammar and lexicon building: some of the grammars and dictionaries of the early 1960s were very large even by current standards.

Research in this period was thoroughly international, with considerable activity in the USSR as well as in the USA and Europe, and some in Japan. US grant funding increased after Sputnik 1, but the work had begun before. Russian and English were the dominant languages, but others, including Chinese, were involved (Booth, 1967; Hutchins, 1986, 2000).

Though the period ended under the cloud of the 1966 ALPAC Report, most of those engaged were neither crooks nor bozos. Many came to NLP research with a background

and established status in linguistic and language study, and were motivated by the belief that something practically useful could be achieved, even though the strategies adopted were crude and the results not of high quality. The first major question was whether even to obtain only limited results, principled methods based on generalisation were required, or whether ad hoc particularisation would suffice. The second issue was the relative emphasis to be placed, in either case, on syntax and on semantics. The third problem was the actual value of the results, especially when balanced against pre- or post- editing requirements.

The main line of work during this period can be summarised as starting with translation as lookup, in dictionary-based word-for-word processing. The need to resolve syntactic and semantic ambiguity, and the former in particular because it is not open to fudging through the use of broad output equivalents, led to ambiguity resolution strategies based on local context, so dictionary entries became in effect individual procedures. Semantic resolution involved both specific word, and semantic category, collocation. But long-distance dependencies, the lack of a transparent word order in languages like German, and also the need for a whole-sentence structure characterisation to obtain properly ordered output, as well as a perceived value in generalisation, led to the development of autonomous sentence grammars and parsers.

Most of the NLP research done in this period was focused on syntax, partly because syntactic processing was manifestly necessary, and partly through implicit or explicit endorsement of the idea of syntax-driven processing. The really new experience in this work, and its contribution to linguistics in general, came from recognising the implications of computing represented by the need not only for an explicit, precise, and complete characterisation of language, but for a well-founded or formal characterisation and, even more importantly, of the need for algorithms to apply this description. Plath's account (1967) of NLP research at Harvard shows this development of computational grammar with its lexicon and parsing strategy very clearly. But as Plath also makes clear, those concentrating on syntax did not suppose that this was all there was to it: the semantic problems and needs of NLP were only too obvious to those aiming, as many MT workers were, at the translation of unrestricted real texts like scientific papers. The strategy was rather to tackle syntax first, if only because semantic ambiguity resolution might be finessed by using words with broad meanings as output because these could be given the necessary more specific interpretations in context.

There were however some workers who concentrated on semantics because they saw it as the really challenging problem, or assumed semantically-driven processing. Thus Masterman's and Ceccato's groups, for example, exploited semantic pattern matching using semantic categories and semantic case frames, and indeed in Ceccato's work (1967) the use of world knowledge to extend linguistic semantics, along with semantic networks as a device for knowledge representation.

MT research was almost killed by the 1966 ALPAC Report, which concluded that MT was nowhere near achievement and led to funding cuts especially in the most active country, the USA, even though it recommended support for computational linguistics. But it is important to recognise what these first NLP workers did achieve. They recognised, and attempted to meet, the requirements of computational language processing, particularly in relation to syntactic analysis, and indeed successfully parsed and characterised sentences. They investigated many aspects of language, like polysemy, and of processing, including generation. They addressed the issues of overall system architectures and processing strategies, for example in direct, interlingual or transfer translation. They began to develop formalisms and tools, and some influential ideas first appeared, like the use of logic for representation (cf Yngve, 1967). Some groups were also established, developing resources like grammars and gaining

experience, as at the Rand Corporation. There was indeed enough knowhow by now for some textbooks, like Hays (1967).

There was little work, on the other hand, on some important problems that have since attracted attention, like anaphor resolution, since though text was being translated it was treated as a sequence of independent sentences, or on the function of language, since the work was mainly on single-source discourse. There was little attempt to incorporate world knowledge, and to relate this non-linguistic to linguistic knowledge, though some world knowledge was smuggled in under the heading of semantics. The belief, or challenge, was that one could get far enough with essentially linguistic, and therefore shallow, processing not involving reasoning on world models. The research of this period did not produce any systems of scope or quality, though by the end of the 1960s there were MT production systems providing output of use to their customers (Hutchins, 1986). There was more merit in the work of the period, and more continuity, through individuals, with later effort, than subsequent myths allow, though the early literature was inaccessible and little used. But perhaps the best comment is Bledsoe's at the International Joint Conference on Artificial Intelligence of 1985 (Bledsoe, 1986) on the value, for artificial intelligence as a whole, of the early MT workers' head-on attempt to do something really hard.

Work on the use of computers for literary and linguistic study also began in this period, but it has never been closely linked with that in NLP, though some common concerns have become more prominent recently.

3 Phase 2: late 1960s to late 1970s

The second phase of NLP work was artificial intelligence (AI)-flavoured, with much more emphasis on world knowledge and on its role in the construction and manipulation of meaning representations. Pioneering work influenced by AI on the problems of addressing and contructing data or knowledge bases began as early as 1961, with the BASEBALL question-answering system (Green et al., 1961). The actual input to these systems was restricted and the language processing involved very simple compared with contemporary MT analysis, but the systems described in Minsky (1968), and Raphael's SIR in particular, recognised and provided for the need for inference on the knowledge base in interpreting and responding to language input.

Woods et al.'s LUNAR (Woods, 1978) and Winograd's SHRDLU (Winograd, 1973) were the natural successors of these systems, but they were widely seen at the time as representing a step up in sophistication, in terms both of their linguistic and their task-processing capabilities. Though differing in many ways they shared a procedural style and were perceived as having an overall coherence as systems and a genuinely computational character. The dominant linguistic theory of the late 1960s, transformational grammar, was seen both as fundamentally unsuited to computation and particularly analysis, even though TG was formally oriented and there was at least one serious transformational parser, and as offering nothing on semantics, which had to be tackled for any actual NLP system. The computational confidence illustrated by Woods' and Winograd's work, and the range of experiment it promoted, while drawing on previous work, is well shown by the varied research reported in Rustin (1973).

The view that current linguistics had nothing to contribute, and the feeling that AI was liberating, were also apparent in Schank's work (1980), which explicitly emphasised

semantics in the form of general-purpose semantics with case structures for representation and semantically-driven processing. The community's concern, illustrated by Winograd and Schank alike, with meaning respresentation and the use of world knowledge then became an argument, reflecting a widespread feeling in AI stimulated by Minsky's promulgation of frames, for the use of a larger scale organisation of knowledge than that represented in NLP by verb case frames or propositional units: this large-scale organisation would characterise the different relationships between the elements of a whole universe of discourse, and would support the inferences, including default inferences, needed especially in interpreting longer discourse and dialogue. NLP would deliver deep representations integrating and filling out individual inputs to form a whole constituting an instantiation of a generic world model. Schank's arguments for the Yale group's use of more event-oriented scripts developed this line in the context of earlier work by linking individual propositional case frames with the larger structures via their semantic primitives (cf Cullingford, 1981). Semantic networks (Bobrow and Collins, 1973; Findler, 1979) were similarly proposed as a third variant on this theme, offering a range of options from associative lexical networks only weakly and implicitly embodying world knowledge to alternative notations for frames. These types of knowledge representation linked NLP with mainstream AI, and their descriptive and functional status, for example in relation to logic, was and has remained a matter for debate.

Semantic primitives seen, as in Schank's Conceptual Dependency, as having a representational and not just a selective role also appeared to fit naturally with the need to capture underlying conceptual relations and identities in discourse processing, particularly for types of material or task where fine distinctions do not figure. Their status too was a matter for controversy, but they have continued in use, supplemented by or sometimes in the form of domain-specific categories, in application systems. They have also had a significant role, in the more conventional form of selectional restrictions, even when semantic driving has been abandoned.

The general confidence of those working in the field, and the widespread belief that progress could be and was being made, was apparent on the one hand in the ARPA Speech Understanding Research (SUR) project (Lea, 1980) and on the other in some major system development projects building database front ends. Several of the SUR projects were ambitious attempts to build genuinely integrated systems combining top-down with bottom-up processing, though unfortunately the best performing system against the target measurements was the least theoretically interesting.

The front end projects (see e.g. Hendrix et al., 1978) were intended to go significantly beyond LUNAR in interfacing to large autonomous (and therefore not controlled) databases, and in being more robust under the pressures of 'ill-formed' input; and the confidence on which they were based drove other work including that on the first significant commercial front end, INTELLECT (Harris, 1984). But these projects unfortunately also showed that even an apparently straightforward, and perhaps the simplest because naturally constrained, NLP task was far more difficult than it seemed to be. NLP workers have been struggling ever since on the one hand with the problems of constructing general-purpose transportable front ends and of providing for the acquisition of application-specific knowledge, and on the other of handling the user's real needs in dialogue. The former led to the development of modular architectures, general-purpose formalisms, and toolkits, typically for supplying a specialised lexicon, semantics, and domain and database model on top of standard syntax, following the sublanguage approach which had been pioneered for text processing by Sager's NYU group (in Kittredge and Lehrberger, 1982), but sometimes supplying a specialised syntax as well.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.