Natural Language Processing in the Undergraduate Curriculum

Robert Dale **Diego Mollá-Aliod** **Rolf Schwitter**

Centre for Language Technology
Macquarie University
Sydney, New South Wales 2109
Email: {rdale,diego,rolfs}@ics.mq.edu.au

Abstract

The paper has two purposes: first, we argue that natural language processing, and particularly those aspects of that field often referred to as language technology, should play an important role in the computer science curriculum; second, we describe in broad terms the content of an undergraduate program we have developed at Macquarie University that covers this material. We question the industrial relevance of much that is taught in NLP courses, and emphasize the need for a practical orientation as a means to growing the size of the field. We argue that a more evangelical approach, both with regard to students and industry, is required. The paper provides an overview of the material we cover, and makes some observations for the future on the basis of our experiences so far.

Keywords: Natural language processing, computational linguistics, language technology.

1 Introduction

Natural language processing (NLP) is the area of study that focuses on techniques that enable machines to work with human language. This involves not only the 'understanding' or analysis of language, but also the generation or production of language. The techniques developed in the field are equally applicable to both spoken and written language, although work that is concerned with the signal processing aspects of speech tends to be considered a relatively separate discipline. Although this is something of a simplification, we can take the view that NLP takes over once the sequence of spoken words has been, in effect, transcribed.

Natural language processing has a long history, going back to early machine translation experiments in the 1950s. It is generally considered to be a subfield of artificial intelligence. A distinction is sometimes drawn between natural language processing and computational linguistics; for some, the latter focuses on more theoretical aspects of the field, although for our present purposes the two terms can be considered synonymous.

In more recent years, the terms 'language technology' and 'human language technology' have been used to refer to work in this area, and particularly work that has a practical import. This 'rebranding' has come about for a variety of reasons, not least that it is now possible to construct commercially viable applications of the technology; these opportunities have in turn been made possible by, on the one hand, the increases in computing power over the last 15 or so years, and on the other, the simultaneous availability of large bodies of text via the World Wide Web. Much current work in NLP and Language Technology focuses on the use of these large corpora of text: in the 1960s, a corpus of one million words would have been considered very large, but these days, billions of words are easily accessible to and manipulable by individual researchers.

Against this background, this paper aims to do two things. First, we argue that natural language processing, and particularly those aspects of that field often referred to as language technology, should play an important role in the computer science curriculum; and second, we describe our experiences in setting up an undergraduate program in the area, with a particular emphasis on the philosophy that lies behind the decisions we have made in designing this program.

In Section 2 we sketch the background to the program, and outline the perspective we take on teaching in this area. Against this backdrop, in Section 3 we describe the orientation and content of the program in some detail. In Section 4 we discuss the evaluation of the program, identify some lessons we have learned regarding what works and what doesn't, and point to where we intend to go in the future.

2 Background

2.1 How The Program Came About

Our program is hosted by the Department of Computing at Macquarie University, which offers a typical range of computer science courses. Standard undergraduate degree programs are three years in length. Students may elect to stay on for a fourth year in order to obtain an honours degree, although in a marketable area like computing, we find that relatively few students stay on beyond third year. The teaching year is split into two thirteen week semesters, and each of our courses is of one semester duration.

In 2000, we obtained Federal Government funding under the Science Lectureships Initiative to set up an undergraduate program in language technology. To obtain this funding, we argued that skills in the language technologies were critical to the development of the next generations of computer interfaces, echoing statements made by many both in industry and academia. Central to our proposal was the identification of the twin streams of (a) spoken language interaction and (b) smart text processing, particularly with regard to the Web; we took the view that these two major areas would define the future of commercial NLP activities over the next five years. Our proposal emphasised heavily a practical orientation, whereby we set our goal to be the training of knowledge workers who will design and develop practical applications in these areas. Our proposal was sup-

ported by a number of industry partners, including the CSIRO and the Australian branches of Motorola, Sun Microsystems, Philips Speech Systems.

2.2 Our Philosophical Orientation

We carried out an informal survey of courses in natural language processing around the world. Our perception was that, in many institutions, natural language processing and computational linguistics courses tended to share two particular characteristics.

First, relatively few institutions have more than one course at undergraduate level that provides material in this area. In many cases, material in NLP or CL appears only as part of a more general course on Artificial Intelligence. This is of course determined by a range of local factors, including inevitably the interests and knowledge of available staff. However, an important factor in many institutions that do not have a long-established and strong research group in the area is the widely-held sentiment that NLP is a somewhat peripheral topic, or a subject of purely theoretical interest. This makes it hard for those staff who are interested in teaching in this area to argue for a significant presence in the curriculum.

A second observation is that the material taught in introductory courses often tends to focus on what we might call computational syntax: writing grammars and building parsers. Again, there are good reasons for this: some would argue that you can't do much else until this material is covered, and this is clearly the corner of NLP that is most well-established with consolidated results, as reflected by the balance of coverage found in texts such as (Allen 1995) and (Jurafsky and Martin 2000), and, perhaps less so than in the past, the topic coverage at conferences such as the Association for Computational Linguistics (ACL) and the International Conference on Computational Linguistics (Coling).[1]

With regard to the first of these observations, we take a strong position. Natural language processing is critical for machine interfaces and information processing technologies of the future. Correspondingly, NLP needs to become a much more central part of computing curricula: every student should be exposed to this area. Our desire, presumably shared by most who work in the area, is to see the field of NLP grow, with many more knowledgeable practitioners, particularly in industry.

With regard to our second observation, however, we take the view that the focus adopted in much undergraduate teaching in this area does not support this goal as well as it might. Teaching students about grammars and parsers may serve as a suitable introduction to further study in the area, but the bulk of students who undertake undergraduate degrees will go on to work in industry; only a minority are likely to work in research laboratories or undertake doctoral studies. Consequently, those graduates who find themselves in a position where they might have the opportunity to use language processing techniques for the development of sophisticated applications are unlikely to have the full range of tools they need at their disposal. The relatively narrow focus of much undergraduate NLP teaching may also be in part responsible for the fairly widespread view amongst those outside the field that NLP is basically about parsing and not much else. This perception results in occasional postings to bulletin boards where senders from outside the NLP research community request a 'parser',

[1] One of the authors recently completed a book project that had as its goal the production of a resource that would meet this concern by providing a more balanced coverage of different aspects of NLP: see Dale et al [2000]. Unfortunately, this book is too large and expensive in its current form for use in our courses.

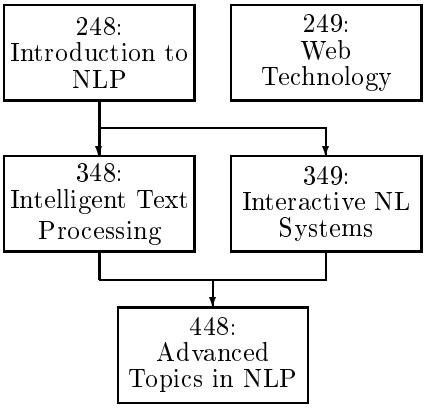

Figure 1: The Prerequisite Structure

with their queries expressed in terms that make it clear that they believe this one component will solve all their NLP problems; those who are familiar with NLP will be aware that a parser is only one component amongst many that is required for a full-blown NLP system.

2.3 The Importance of the Job Market

We believe that if NLP is really to grow into a field of substantial visibility and worth in the wider industry community, there is a need to raise the status of study in NLP beyond that of a niche interest. The key to making this happen is to emphasize the practical utility of work in the field.

There is a real chicken-and-egg situation here. We will only see an explosion in the number of real NLP applications if there are more well-rounded NLP practitioners working in industry exploring and developing those applications; but students are very savvy about the job market, and, faced with a choice, are unlikely to choose an NLP course over, say, a networking course, when faced with the relative proportions of job ads they see in the press and on the Web.

There are two related consequences of this. First, evangelism is critical: we need to get more trained students out there, offering NLP solutions to problems. At the same time, we need to give students concepts and techniques that enable them to provide those solutions. We need to provide material that students can see is relevant, and that can be used in many contexts.

In our analysis, the job market for skills in language processing, to the extent that it is identifiable, consists of two major segments. First, and most obviously, there are companies that develop voice applications: there are a great many companies now working in this area, and voice recognition is a recognized industry sector. Second, there are companies that might use NLP techniques in developing applications that process, maintain and reuse documents, whether on the desktop or on the Web. While the first of these segments is quite clearly identifiable, it is much more difficult to identify a sector that focuses on using NLP techniques on text. With some notable exceptions (and these are largely small startups), we do not tend to find companies whose focus is NLP. This is not really surprising; NLP is just one tool amongst many that might be used in document processing, and document processing is something that manifests itself in many contexts.

We therefore have a particular challenge here: we need to communicate to students that NLP is something they may be able to use in their future careers,

but we can't point to many job ads that specifically request NLP skills. The intuition of those working in the field is that this stuff ought to be something that can make a difference in the processing of documents, but there is not a lot of visible evidence that it is being used in those situations. Anecdotal personal experience suggests that many companies would benefit from the application of NLP skills but are not aware of this. One suspects that organizations may often be making use of techniques that we might want to think of as NLP, but that these techniques are not recognized as such.

3 The Program

Given the above, our goal was to construct a range of courses that covered a broad range of material that students might be able to use in their subsequent careers. To emphasise the practical orientation of what we wanted to do, we deliberately pitched the program as being concerned with Language Technology, rather than as a program in either Natural Language Processing or Computational Linguistics.

There is clearly something of an evangelical element to this: we wanted to make students aware of a broad range of techniques that we would label Language Technology, with the goal that, over time and as these students enter the work force, an awareness would start to spread that these techniques are widely usable. This is not a short-term strategy: it takes several years for the results of these efforts to permeate through the system to a stage where they can be evaluated, but it is essential to get started.

In this section, we present a summary of the material we deliver in the courses that make up our program. More detail on each of these courses, and the program as a whole, can be found at `http://www.clt.mq.edu.au/Teaching`. The program consists of four courses that focus principally on Language Technology, and an additional course that looks more broadly at technologies for working with the Web. Figure 1 shows the prerequisite structure that currently holds between these courses.

3.1 Comp248: An Introduction to Natural Language Processing

Taught in the second half of second year, this is the course in our program that most closely matches the typical undergraduate NLP course. The design of this course was driven by a desire to show students that they could build a useful, functioning application using NLP techniques; to this end, we felt it was important not to teach only computational syntax, but also something about semantics. Our position here is that syntactic processing is only a means to an end, and we felt it important to quickly get students to the stage where they could actually see some practical import of what they were doing. To this end, in the first half of the course we take a fairly standard approach to teaching Prolog, whereby the students do some rudimentary morphological processing, build some Definite Clause Grammars, and learn about parsing techniques. In the second half of the course, we add semantics to the mix: although we teach an introduction to lambda calculus at this stage, for the practical work we focus on a much shallower approach to semantics (effectively semantic grammars), and the students build a NL database query system that allows them to ask questions of a database of flights. Along the way they learn about unification-based grammar, case frames, lexical resources, WordNet, and semantic networks. The guiding principle throughout is relevance to building a practical application.

3.2 Comp249: Web Technology

Although this course is part of our Language Technology program, it does not contain a significant language technology element (at least as the term is currently construed). It turns out that the background material taught here has proven to be very useful in other courses we teach, so we are considering binding this course more tightly to the others. The course covers: Perl programming, Web design, client-server computing, search engines, XML and related technologies, database integration, privacy and security, VoiceXML, and content management; inevitably, with such broad coverage, most topics are treated relatively briefly.

Our goal for this course is to target a student body who have little awareness of what NLP is and to get them to see LT in a wider perspective. The success of this course, which is by far the most popular of the units in the program, has led us to explore better ways of leveraging this interest.

3.3 Comp348: Intelligent Text Processing

At the third year level, we offer two courses that take the second year material as a base. We noted earlier that we viewed the job market as consisting of two relatively distinct sectors, one concerned with voice processing and one concerned with document processing. This perception is very deliberately reflected in the individual biases of the third year offerings; Comp348 addresses the needs of document processing, whereas Comp349, discussed later, leans more towards voice processing.

The course on intelligent text processing covers basics of text processing using Perl; tokenisation and sentence segmentation; text summarisation; information retrieval; corpus-based approaches; part of speech tagging; word sense disambiguation; information extraction; and machine translation. Again, this is a lot of material to cover, and inevitably we only skim the surface of many topics. However, in the first offering of the course, students did significant assignments in both text summarisation (using sentence extraction) and information extraction. The latter assignment was run roughly along the lines of the Message Understanding Conferences: using conference announcements as a data set, the students were provided with a training set on the basis of which they built an information extraction system. This was then tested against unseen data, and scores were automatically derived. Now in its second offering, our intention is to use anaphor resoluton as the focus of an assignment.

Our goal in this course is to provide students with a toolset for text processing from a language technology perspective. We focus on relatively shallow methods, since these are the methods students are most likely to find themselves using in their subsequent careers. Our driving aim here is for our alumni to recognize that LT provides solutions.

3.4 Comp349: Interactive Natural Language Systems

As already indicated, this course aims to provide knowledge that students need in order to be effective in the voice processing industry sector.

The focus here is on, effectively, text- and speech-based dialog systems. In the first half of the course, we cover a significant amount of relatively theoretical material, covering question answering systems, database interfaces, and answer extraction. Students build a quite sophisticated text-based natural language query system.

In the second half of the course, we attempt to apply the theoretical ideas in the very practical context of building spoken language dialog systems. We begin by using the CSLU Toolkit,[2] which the students use to build a voice banking application. We then introduce VoiceXML in some detail; using a PC-based development environment, students build a simple flight reservations system.

We place a heavy emphasis here on aspects of voice user inferface (VUI) design; in the practical half of the course, the materials we use take a similar approach to that taken in vendor courses that aim to train dialog designers and grammar writers. At the same time, we have as an important aim a clear exposition of the relationship between the ideas explored in research systems and commercially deployed systems; in practice it can be very hard to see a path from the former to the latter. We make clear to students that our goal is to teach them how to build practical dialog applications now, but to get them to think about what the next generations of such applications might be in the light of the results that come out of research laboratories.

3.5 Comp448: Advanced Topics in Natural Language Processing

For those students who stay on for a fourth year, we run a course that is more driven by a selection of specific research topics. We are using this course to cover in more depth core topics that are only really touched upon in earlier courses, with more detailed exploration of word sense disambiguation, anaphora resolution, discourse structure and natural language generation. The course is seminar-based, with a high proportion of student presentations, and an assignment in anaphor resolution.

The level of interest amongst students at this level was such that, in the year we first offered the honours level course, we also ended up offering two additional honours level courses in the language technology area, one on speech recognition and one on question answering systems.

4 Outcomes and Issues

The program has been operating since the second half of 2000. Since that time, we have taught Comp248 three times; Comp249, Comp348 and Comp349 twice; and Comp448 once.

It is too early to establish to what extent the material we have taught is impacting on graduates' work practices: the first students to complete degrees that incorporate our courses have only recently graduated. However, we have made use of a number of feedback and review mechanisms over the last 18 months, and these have already provided us new ideas for how to improve what we are trying to do.

4.1 Evaluating Course Content

We make use of the typical infrastructure made available for evaluation purposes: student-staff liaison committees, formal questionnaires, and also a significant amount of informal feedback through discussions with students. We also have a management advisory board with representation from industry; this meets twice a year to review the development of the program and to comment on its industrial relevance. At this early stage in the development of the program, we feel that it is very important to scrutinize carefully what we are doing; accordingly, we carry out a formal review of each course on completion of an offering, with the results feeding back into sometimes substantial revisions of the material for subsequent offerings.

Generally, the courses have been extremely well received by the students who take them: not infrequently we have had students comment that these are the most interesting courses they have taken as part of their degree. Consequently, we have strong evidence that students find the material interesting, challenging and informative. Our advisory board is very comfortable with the material we teach, but we suffer here from the problem that the voice recognition industry is better represented here than the hard-to-define document processing industry alluded to earlier. Our industry partners think we are going in the right direction; but we have yet to demonstrate that the wider industry community will see a benefit from students who have grasped this material.

4.2 Course Materials

We have faced a not insignificant problem in finding appropriate course materials for these courses, with the consequence that we have had to develop most things from scratch. For the first offering of Comp248, the introductory NLP course, we used Allen (1995); in the second and third offerings, we found Covington (1994) to be more useful. Although this is technically out of print, Prentice Hall has a technology for producing short print runs on demand.

The materials problem was more severe in our third year courses, since there are no even vaguely adequate textbooks for the material we wanted to cover. We provide students with a comprehensive reading packet, but it is not easy to find appropriate survey or introductory readings in the various topic areas we cover. As a consequence of this we are exploring the possibility of writing a textbook that covers the material in each of these courses.

5 Lessons Learned and Future Directions

Two years from the start of the program, we are reasonably assured that we are going in the right direction; some things, inevitably, require fine tuning. We note here some key consequences of our experiences so far.

5.1 Voice Captures the Imagination

Perhaps not surprisingly, it is the study of voice recognition that has really captured students' imaginations. The level of enthusiasm generated in a laboratory full of students wearing headsets talking to their machines is wonderful to watch (although the working environment doesn't do a lot for speech recognizer accuracy). With this in mind, the most recent offering of our second year course, Comp248, contains some of the voice material previously used in the third year Comp349 course. We are also incorporating an emphasis here on technology that students might meet outside of the curriculum, such as chatterbots. Our strategy here is to entice students into the area with appealing content, and draw them into the more theoretically challenging material in later courses.

5.2 Document Processing as a Theme

It has become obvious that our Web Technology course could play a more coherent role in our program. One obvious direction we are pursuing is to

[2] This toolkit provides an excellent environment for teaching students to think about issues such as dialog flow, as well as introducing them to many other aspects of spoken language dialog systems. See http:// cslu.cse.ogi.edu/toolkit/.

cement the two strands identified earlier even further, by seeing the Web Technology course specifically as a precursor for the Intelligent Text Processing course. At the same time, we are considering broadening the third year course to cover Document Processing more generally, as a way of making its relevance more apparent.

5.3 Linguistic Background

We have met the common, and not unexpected, problem that some students do not have a sufficient grasp of linguistic matters to perform satisfactorily in this area. To this end, we have initiated the introduction of a first year course that covers basic aspects of linguistics, logic and computation, taught by ourselves in conjunction with the University's Departments of Philosophy and Linguistics.

5.4 Conclusions

So far, our program has been seen as very successful from an academic perspective, and has generated significant interest amongst students. Our next challenge is to persuade industry to see students with this training as very valuable assets. We have instituted an alumni program that will attempt to track these students, with the expectation of some preliminary feedback being available by mid 2003.

References

Allen, J. (1995), *Natural Language Understanding*, Benjamin Cummings, Menlo Park, CA.

Covington, M. (1994), *Natural Language Processing for Prolog Programmers*, Prentice Hall, NJ.

Dale, R; Moisl, H; and Somers, H (eds). (2000), *Handbook of Natural Language Processing*, Marcel Dekker, New York.

Jurafsky, D. & Martin, J. (2000)*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Prentice Hall, NJ.