Planning the future of natural language research (even in Canada)

Graeme Hirst
Department of Computer Science
University of Toronto

May 1990

For this talk, I've tried to include something to offend everybody. Um, if for any reason, you're not offended when I finish, then see me afterwards and I'll say something insulting — especially for you.

1. Introduction

What I want to talk about is:

- What research in NL has done — and is likely to do.
- What it could do for Canada.
- Why it hasn't done it — yet.
- What we need for it to happen.

Like other parts of AI, Canadian NL research has made significant contributions to the field. But many of our best people have left the field, or gone south. NL doesn't have the small but distinguished nucleus in Canada that some other AI subfields have.

An invited talk presented at the Eighth Canadian Conference on Artificial Intelligence, Ottawa, 24 May 1990.

Like other parts of AI, Canadian NL research has been relatively lucky compared to many other sciences. AI is still a ‘‘glamour field’’, even if not as glamorous as putting a Canadian in space. But NL hasn’t been quite as lucky — it’s had little support from CIAR, for example. (I won’t speculate on the cause/effect relationship between this and the lack of a distinguished nucleus.)

That’s all rather unfortunate. NL research could be important for Canada, as it already is in Europe (a point I’ll come back to shortly) — perhaps even more important than putting a Canadian in space.

2. Where NLP is now

Let me outline the current state of natural language processing research in the world.

I’d better start by defining the area that I’m going to cover, with an apology for the terminology. ‘‘Natural language processing’’, or ‘‘NLP’’, or ‘‘computational linguistics’’, is the subfield of AI that’s concerned with the use of human languages — natural languages. That means, in particular, building systems that can deal with the structure and content of language the way a human would — not just processing language the way a word processor, say, does. The field is sometimes called ‘‘NLU’’, ‘‘natural language understanding’’, but I’ll avoid that term because I don’t want to limit myself, or the field, just to the *comprehension* of language.

NLP, perhaps even more than other parts of AI, is an incremental science. That is, it’s not given to breakthroughs or elegant theorems; rather, it is an accumulation of smaller ideas, techniques, and formalizations that together build up a system.

This reflects the nature of language itself. Language has evolved to take advantage of many different parts of human cognition in its operation. Language understanding draws on many different kinds of knowledge. It seems to use both hardware in the brain that's specialized for language and also the brain's more general-purpose hardware. Language pervades cognition. So it's hardly a surprise that language understanding programs often look like an agglomeration of bits and pieces.

NLP has come a long way since the 50s and 60s — thanks in part to the co-evolution of modern linguistics. Not only has the development of modern syntactic and semantic theories strongly influenced NL research, but the concerns of NLP for process-oriented theories of language have had a strong influence on theoretical linguistics. The two fields are now closer than ever before.

The same is true of psycholinguistics. Computational work on NLP has influenced the development of psycholinguistic models of how people process language, and experimental work on how people understand has influenced the development of techniques in NLP.

As an example of where the field is today, consider parsing. It's now clear that, contrary to some suggestions in the 70s, a complete syntactic analysis of a sentence *is* necessary as a preliminary to semantic analysis. And we now have fairly large, robust systems for parsing — not just toys — that are, for example, used in corpus studies in which large bodies of text are parsed or processed.

Semantic analysis is also doing well. We have new approaches to semantic interpretation, to knowledge-based ambiguity resolution, and to fitting compositional semantics into unification-based systems.

We are rapidly learning to use on-line dictionaries and reference works. There's an awful lot of useful knowledge in such books. They were originally intended for use by people, but methods that enable NLP programs to use them are a matter of current research.

Language generation and user modelling are starting to come into their own. These used to be just small areas within NLP.

The importance of work in language generation is obvious. For one thing, it's the second half of a machine translation system. And as on-line knowledge bases get bigger and more complex, we need a way for their contents to be expressed to humans. This research has got to the stage of considering the production of whole paragraphs. Given a pile of ideas, represented in some knowledge representation formalism, which the system has to ‘‘say’’, there are problems in deciding what order to say the ideas; how to break them into sentences; how to get the emphasis on the more important parts; and how to make the whole result sound natural. This is a complex task in planning, and there's a lot of interesting work going on applying planning techniques to language generation.

User modelling is likewise important. In many applications, especially, I think, in education, an NLP system can't just treat its user as a generic human being, but has to take into account their level of knowledge, their goals, and, of particular interest, their beliefs and possible misconceptions or misunderstandings.

One area we're *not* so far along in is interpretation in serious domains. There's still considerable work needed in NL-oriented KR, ontology building, etc. The problem is that no present-day logic or knowledge representation formalism can express everything that NL can, and that rather puts a damper on efforts to interpret NL into such formalisms. There is some promising work in the area, um, my own, for example, but not enough attention is being paid yet to the needs here.

In general, there is a greater emphasis in the field on ‘‘real’’ language use. That is, we are worried about actual texts written by actual people, and not just artificial examples like ‘‘John saw the spy with the telescope’’.

Applications are filtering into the world:

- Grammar checkers, although presently modest, are starting to appear for personal systems.

At present, they're still pretty dumb, but there's no doubt that they'll get better.

- NL database interfaces, both for large systems and for personal systems, are now widely available.

Symantec's Q&A system is one of the best known. Incidentally, it's a direct spin-off from NLP research done in the 70s by the NL group at SRI International in California.

In general, we can expect to see much greater use of NL in information systems — both for communication with the systems, and in the knowledge base of the systems themselves.

- News- and message-routing systems, which use some understanding of the content of the item to decide where to send it, are in regular production use.

These are systems that, for example, read newswire stories as they come in and classify them to be sent to whoever needs them. Such systems are now used in financial houses to keep analysts aware of what's happening that may be relevant to their investment decisions. Reuters uses them to index its news stories. And the U.S. military wants to use them to process certain kinds of operation messages.

- There are many machine-aided translation systems, some quite sophisticated, now available.

These systems, of course, are not fully automatic but work together with a human translator. I think in a few years, we'll start seeing systems that can be assisted by the unilingual writer of a document, rather than needing a professional translator. For example, an English-speaker could write a memo or an e-mail message, and then, knowing the content of the message, help the system translate it into French or Japanese — without having to know either of the target languages. I'll come back to MT later in the talk.

- Intelligent computer-assisted language instruction is lagging behind for some reason — possibly because it's not yet understood how to use them pedagogically — but there's some interesting work in the area.

In general, NL research applications have use in just about any domain or any industry that uses language or information systems, and that's just about everywhere and anywhere that communication takes place.

3. What can NLP research do for Canada?

Well, now, what are the potential benefits of NLP research for Canada? Let’s look at what’s happening in other countries.

Although the basic research is largely centred in the U.S., much of the action in applied research and development is in Europe and Japan.

Europeans have always been more conscious of language than North Americans — even Canadians. Any educated European is assumed to speak two or three languages. (Interestingly, the exceptions to that seem to be the English and the French.) Many European countries are bilingual or multi-lingual — Belgium, Switzerland. Many have languages that are spoken by few outsiders — the Netherlands, Denmark, most central European countries — so it is essential to learn the more widely spoken languages. And the EC, which has nine official languages, and the single European market coming in 1992, are making an emphasis on language even more important.

Thus Europe has what are called ‘‘language industries’’, while we in North America hardly even have the *term*. There’s an emphasis on translation, and computerization of all aspects of language — writing, translating, managing multi-lingual documentation. The EC is supporting Eurotra, a huge MT project.

Likewise, in Japan. The Japanese Fifth Generation project includes a large amount of NL research. Enormous effort is going into machine translation. And the Japanese Electronic Dictionary project is, I understand, more far-reaching than any comparable Western project. In fact, Marshall Unger has argued in his book that the *real* motivation of the project is not international economic dominance. Rather, it's simply to deal with the domestic problems caused by the complexities for computers of the Japanese language and writing system. And language also features prominently in the applications proposed for the successor project in massively parallel computing.

4. What could have happened in Canada?

4.1. The relevance of NL research to Canada

This is all rather relevant to Canada.

We also have a large need for translation in Canada. And it's expensive and there's always a shortage of qualified translators.

We have a great need for teaching languages in Canada. And not just teaching a second official language, but also, for many immigrants, a first official language.

Anything that improves translation or language teaching is surely of benefit to Canada — both economically and socially. Anything that can make it easier, anything that can make it cheaper.

And, on a larger scale, language is central to the future economy of Canada. We keep hearing about how Canada has to lower its dependence on a resource-based economy. About how the so much of the world economy will be — is already! — based on information and services. About how the typical worker of the future will be a ‘‘knowledge worker’’.

Well, language is how people represent knowledge.

So if computers and automation are to be involved in this knowledge work — and I think we'd all agree that for efficiency they must be — then it would be a good idea if the computers involved could process language. And process it not just as text, like a word processor does, without regard to its meaning, but process it as a repository of knowledge.

Even in artificial intelligence, people seem to have lost sight of this. People rightly see AI as having great potential, and are willing to spend time and money on work in expert systems and formalisms for knowledge representation and reasoning. But what they've forgotten is that knowledge comes from people, and is for the benefit of people. So we need to also be concerned with people's knowledge representation formalisms — natural language — as well as those for computers, and we need to worry about mapping between the two kinds of representation.

After all, there's a vast amount of knowledge out there in the world. About 10 to the minus 87 percent of it is presently in a form suitable for use in any AI system. Of the other 99.9999 percent, a good slab, maybe half, is in natural language. The rest is in people's heads in a non-linguistic form, and when it comes out it does so in the form of actions or, again, language.

This is tacitly recognized in the recent emergence of an AI subfield called ‘‘knowledge acquisition’’. Knowledge acquisition studies methods for copying knowledge from heads into programs. In a sense, the problem of knowledge acquisition is just another form of the general problem of language understanding — a particularly difficult form, because it typically involves complex ideas in complex language.

4.2. Not in Canada? Pity!

So for all these reasons, NLP research could be of great benefit to Canada.

What's more, we could have been a leader!

The TAUM METEO project (at the University of Montreal, in the 70s) was a world leader in MT. I've found that the project and the people who were on it still command enormous respect in the MT community — outside Canada, that is.

In the latter part of the 1970s, Canada was almost unique in having good AI people, good MT people, and, as it happens, a significant number of good people working in computer applications in the linguistic humanities. With a base of such people working together,

- Canada could have been a leader in MT;
- Canada could have been a leader in NLP in general;
- Canada could have been a leader in ICALI;
- Canada could have been a leader in multi-lingual processing.

Canada isn't any of those things. Why?

Funding was withdrawn from the TAUM group because MT wasn't found to be immediately cost-effective in the short term! How incredibly short-sighted — and typically Canadian. We had something good and we blew it. What should have been thought of as basic research was evaluated as if it were product development.

And, in general, Canada couldn't — or at least didn't — match the resources and opportunities available to researchers in the south.

5. Current work

So what are we doing in NL research now in Canada? I can't be a spokesperson for other groups, but I'll briefly mention some of the projects we have worked on at Toronto in the past five years.

- What we've called ‘‘theoretical MT’’:

We don't have the resources to work on real MT projects, but some of our theoretical work has been explicitly directed towards developing ideas that could be used by someone else in machine translation systems.

For example, we've been looking at the problem of *style* in language. There are always many different ways to say the same thing, but they generally differ in subtle but important ways. A translation, if it's to be faithful, has to preserve the nuances of the original text — though the ways in which those nuances may be actually realized in the text could be very different in the two languages involved.

I should emphasize here that I'm not talking about *literary* style, but rather the stylistic questions that arise in ordinary, everyday text like newspapers or computer manuals.

Chrysanne DiMarco recently completed a dissertation on computational formalisms for describing an author’s stylistic intent so that a machine translation system can preserve linguistic style in its translation. The formalism describes style at three levels of abstraction, and only the bottom level is language-dependent. DiMarco has built a system that analyzes the style of input sentences in English or French. And our student, Mark Ryan, has looked at the relationship between style and the linguistic focus of a text. We hope that this work will develop into a complete stylistic component for an MT system. I think I can say that this work is good and important and nothing like it is being done almost anywhere else.

A related thesis by Mara Miezitis concerned the problem of lexical choice in translation: How to organize a lexicon so as to know (without exhaustive search) what words the target language makes available to express the ideas in the input language text. Language tends to be very capricious in this area. For example, English has a word for someone in their eighties, ‘‘octogenarian’’, and another for someone in between 13 and 19, ‘‘teenager’’. But there’s no word for someone in their forties, and you have construct a phrase such as that one, ‘‘someone in their forties’’. When you’re translating between languages, you need to know whether the language you’re translating into has a single word or idiomatic phrase for some parcel of concepts, or whether a phrase has to be constructed from words representing components of that parcel. And you need to be able to do that without exhaustively searching your dictionary. Miezitis’s method is, in effect, a clever way of organizing and indexing the lexicon by concept in order to make the search efficient.

- Prototype ICALI systems:

Again, our group doesn’t have the resources to be building large, complete CAI systems, let alone developing courseware and evaluating the systems with real language learners. So again we have concentrated on theoretical work directed toward application in such systems.

One application is in teaching grammar to students who are starting to learn a second language. It would be nice if the computer could parse the students’ sentences and tell them if they were right or wrong, and if wrong then exactly what the error is. However, a regular parser can’t do that, because if you give a regular parser an ungrammatical string it will eventually fail, but the point of failure need be nowhere near the actual error. Nor is it always obvious exactly what the error is. For example, two parts of a sentence may simply not work together, but that doesn’t tell you which, if either, is quote-wrong-unquote.

Mark Catt has devised a parser for use in a language teaching system that instead of just failing on ungrammatical input can diagnose where the error is and give feedback to the student. Moreover, many errors that learners make come from wrongly trying to apply rules from their native language to the language that they’re learning. Catt’s parser can take the student’s native language into account in determining the error that was made.

Julie Payette is now applying Catt’s parser and some aspects of Chrysanne DiMarco’s work on style to develop a system for teaching nuances of language to advanced learners.

- Knowledge acquisition as a problem in language understanding:

Stephen Regoczei (from Trent University) and I have been considering what's involved in knowledge acquisition for knowledge-based systems just as a manual task in which an analyst — or ‘‘knowledge engineer’’ as we say these days — has to convert an expert's utterances into some formal knowledge representation. We've concentrated on the idea that understanding consists, in effect, of the understander *adding* his or her or its existing knowledge to the text — what we've called ‘‘concept-cluster attachment’’.

- Knowledge representation for language understanding:

I mentioned earlier the problem that no known logic or KR formalism can express everything that NL can. One particular problem that I've been looking at is assertions of non-existence, as in sentences like ‘‘The lecture was cancelled’’ and ‘‘The strike was averted’’. In most KR formalisms, just to use a term is to assert that its denotation exists, and that's no good if you're trying to assert the opposite. An obvious solution is to introduce an ‘existence’ predicate. But unfortunately, as philosophers have known for the last 200 years, that gets you into all sorts of trouble. My solution has been to steer around those problems by augmenting an analysis by Terence Parsons, a contemporary philosopher, with a multi-faceted view of types of existence.

- And besides all this, we've also been doing basic research in various other aspects of NL.

These projects are in areas such as semantics, ambiguity resolution, and linguistic pragmatics; and I'd be happy to talk further about them afterwards.

6. Prospects for the future

Well, so far I’ve told you that NL research is steaming ahead in the rest of the world, while Canada missed the boat. But some of us are still pushing ahead, rowing as hard as we can. The question is, what we’d need in Canada in order to catch up with the boat.

What we need are several NL projects, at universities or elsewhere, that would be big enough to have an effect: to get researchers together, to do pre-competitive research (as it’s called these days), to show what can be done. With enough infrastructure to get the job done properly, to train graduate students, and to give the students somewhere to work when they graduate.

I’m thinking of projects comparable to, say, the Center for Machine Translation at Carnegie Mellon. The Center has long-term goals in MT applications, and pursues both those applications and basic research directed toward them. It has external funding, provides a place for grad students to learn and work, and has an active visitors program to promote the exchange of ideas. Canada, unfortunately, has little tradition of large project centres like this at universities.

Ideally, we’d want several such projects. We don’t want to get the whole country committed to just one group’s paradigm or one group’s approach.

What are the obstacles to realizing this?

- A lack of money.

Large projects need not just researchers, but space, equipment, programmers, assistants. Building a ‘‘real’’ MT system is 90% software hackwork.

- A lack of momentum, a long up-to-speed time, and already being quite a way behind.
- A lack of industrial and government interest.

We all know the sad story of the state of R&D in Canada. It’s so well known that even Maclean’s had an article on it last week — ‘‘A critical science gap; Canada’s spending on research is falling far behind that of its competitors’’ — with the obligatory picture of the Canadarm, as if that’s the only thing Canadian science has ever produced. We spend a much smaller fraction of our GNP on R&D than other industrialized nations. And the present government has repeatedly broken its promises to change that.

In Ontario and elsewhere, university funds have been cut.

The NRC itself has a doubtful future.

The federal and Ontario governments and the CIAR have initiated their respective ‘‘centres of excellence’’ programs — which are better than nothing, and NL research has had a little benefit from them. But by their nature, they are an admission of failure of the research funding system. And neither ‘‘centres of excellence’’ programs nor NSERC strategic grants, as they presently work, are intended for the kind of bootstrapping that I’m talking about here of a research field that’s at the pre-pre-competitive stage. Centre-of-excellence programs imply existing momentum and quoteexcellence-unquote; strategic grants imply existing industrial interest.

And even where there is some interest in AI, there seems to be little appreciation of NLP research.

Could it still happen? Could we develop NLP research in Canada?

What are our resources?

- A few NLP people in universities and elsewhere. And many more who could be repatriated, including those who have moved to other sub-fields of AI. Good people in other areas of AI and in computing in the linguistic humanities. And lots of keen, good students.
- A few companies that could apply some of the research.
- Larger R&D companies that might be induced to support long-term research, like companies such as Bell Labs and Bellcore do in the U.S. And perhaps CIAR might be too.

But, what else do we need?

- We, as a natural language community, need to effectively communicate our visions and our commitment.
- Active interest from government and industry to support this vision and commitment: A realization that there could be long-term economic benefit from support of NLP research.
- Money.

But we have to be wary of the problem of inadequate half-measures. That is, being offered just a little money, and then written off when we fail to perform miracles on a shoestring.

7. Conclusion

You might wonder why I'm telling you all this. It sounds a bit like a grant proposal. Perhaps I just want a sympathetic ear. I think we're doing some good NL research in Canada, but we're not getting the resources we need.

I'll finish up at this point so there'll be plenty of time for questions.

My main points are that NLP is a viable subfield of AI, and one that could have special importance to Canada — perhaps even as important as putting a Canadian in space. But if it's to realize its potential, it's going to need greater recognition.

Thank you.

Questions

I'll take some questions now, and I'm going to use my prerogative as speaker to ask the first few questions myself.

Q1. Professor Hirst, you didn't mention the word ‘‘GigaText’’ in your talk at all. Comments?

A1. GigaText is AI's own Sprung greenhouse! Since it happened in Saskatchewan, the details aren't well known elsewhere, which doesn't say much for the Canadian media, but which is fortunate for us because GigaText has certainly damaged the credibility of MT.

For those of you who don't know the details: Saskatchewan needed to translate statutes into French. Two Canadian AI people approached the government and promised, in return for money, to set up a MT company using some fabulous new ideas about MT that they had, translate the statutes, and put Saskatchewan on the cutting edge of MT research. The government handed over the money without much thought, and last year the company, GigaText, went broke without delivering on its promises. The Saskatchewan government lost its money, and it has been suggested that GigaText did not use the money in the most prudent and economical way.

Now the Saskatchewan government may have had bad technical advice and been led to have completely unrealistic expectations, but at least they had a vision!

In contrast to GigaText, what I'd like to see is an establishment that will perform basic and long-term research, and apply results. But it's very hard to predict what's doable when. We can't take the approach of ‘‘scheduled space shots’’, or translate 40 statutes automatically by next June, no matter what politicians would like.

Q2. You were critical of the notion of special programmes giving funding for ‘‘centres of excellence’’. Comments?

A2. Centres-of-excellence programmes are like food banks. Almost by definition, they are an admission that our research-funding system has failed! — Not supporting adequately the best people in the field, so a special band-aid programme is needed to bring their resources up to (near) adequacy — to what a proper system would have given them in the first place.

The U.S. doesn't need such programmes! To the extent centres of excellence exist in the U.S., e.g., MIT, Stanford, they tend to be automatic by-products of the system.

Now, such programmes are certainly welcome as better than nothing. But for each Centre of Excellence we need half-a-dozen Centres of Okayness. A research field doesn't move just by supporting the few superstars. They can't do all the work by themselves; other competent researchers need adequate funding too.

Q3. It's not all sweetness and light in the U.S., you know.

A3. That's true. In fairness, let me say what we have in Canada that's better than in the U.S.

The main thing is the way NSERC operating grants work. They permit research that is open-ended and curiosity-driven, not project-oriented. The money (what little NSERC has to distribute) is ‘‘more real’’ than that from U.S. grants because no overheads, PI salaries etc, are taken out; but of course, that means that those costs have to be paid from other funds. NSERC grantees are not dependent on making their work suit the military or government policy fashions the way DARPA grantees are in the

U.S. And with NSERC, one is able to spend more time on the research and less just writing grant proposals.

The only other advantages we have in Canada seem to be cleaner streets and a general moral superiority . . . of some kind.

Q4. Eurotra is not doing so well either, you know.

A4. So I understand. But the difficulties seem to be related to its multi-site structure and to national and scientific rivalries rather than the basic idea of such a project.

Q5. Instead of setting up any kind of rival project, perhaps Canada or Canadian researchers should ask to be included in Eurotra or other projects?

A5. That might be nice. But we’d have to have something special that they would want enough to be willing to make us a co-developer rather than a customer. It’s not clear that we do. And I would imagine that participation in Eurotra would require some pretty special high-level agreements between the Canadian government and the EC.

Q6. I think you’ve been unnecessarily hard on the government. While it’s true that our R&D spending is low, the government’s share is nearly as big as that of other countries. The shortfall is due to industry not doing its part.

A6. That’s true. But why isn’t industry doing its part? Because it has little incentive to do so. Because research is done at companies’ head offices in the U.S. Because there’s no tradition of any need for much research in what was once a more resource-based economy — just dig it up or cut it down and ship it out. Because we don’t have the tradition of venture capitalists willing to gamble on small leading-edge research-based companies.

It *is* the government’s job to try to change this. After the debacle of the Scientific Research Tax Credit program, one can understand a reluctance to try something new. But they did promise that they would. And they haven’t.

Q7. You say you are doing good ‘‘theoretical MT’’ work. Perhaps that’s what our contribution to the world should be. We can do this sort of thing well exactly because we aren’t distracted by large projects requiring real deliverables!

A7. Well, even if that’s true, it would still be nice to have a bit more recognition and support for the work. If Canadian government policy is to be simply that Canadian research, to the extent it exists at all, is our altruistic gift to the world of science for others to exploit at will, then what can I say? It seems to me that perhaps we can be just as altruistic *and* benefit more ourselves.