Last Words

On Becoming a Discipline

Mark Steedman*
University of Edinburgh

The title of this column, Last Words, reminds me of an occasion in 2005, when I had the privilege of attending the award ceremony for the prestigious Benjamin Franklin Medal, given annually to a few scientists who have made outstanding lifetime contributions to science. This time, a computational linguist, Aravind Joshi, was among them, so several past, present, and future presidents and officers of the ACL joined the Great and the Good at the ceremony at the Franklin Institute in Philadelphia.

The eight medal recipients were each represented by a short video presentation, which mostly consisted of voice-over by a narrator, interspersed with sound-bites from the recipients about their life and work, in the last of which they had clearly been asked to deliver as their last words a brief take-home message.

I couldn't help noticing that the warmest applause was reserved for the physicist, a distinguished pioneer of string theory. I was initially puzzled by the enthusiasm on the part of a mostly lay audience for such theoretical work, which for all its elegance and beauty, could not (as far as I could see) be expected to have nearly as much impact on their everyday lives as that of some of the other recipients, who that year included not only Aravind, but another computer scientist whose impact on information processing will be obvious to the members of ACL, Andrew Viterbi.

But then I recalled that the physicist's take-home message had had nothing to do with string theory. This admirable man's last words to us had been the following:

> Everything is made of particles. *So physics is very important*.

1. The Public Image of a Science

I realized then that what we were applauding was not a physicist, or the beauty of string theory, but *physics itself*. I was reminded that physicists are looked on as public intellectuals who can be relied on to deliver Truth, and who are generally assumed to be doing A Good Job, even when they say that the universe must really be ten-dimensional, except that exactly six of the dimensions are curled up so tightly that there is no conceivable way of detecting them, nor any prospect of bringing to bear the huge energies that would be required to straighten them out a bit so we could take a look at them. (As many of you will know, I Am Not Making This Up.)

* School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh UK EH8 9LW.
E-mail: steedman@inf.ed.ac.uk. These remarks are excerpted from the author's Presidential Address to the 45th Annual Meeting of the ACL, Prague, June 2007. I'm grateful to that audience; to Mona Diab, Julia Hockenmaier, Miles Osborne, and Bonnie Webber for comments and advice; and to Franz Och for Google Language Tools.

How different this public image is from that of computational linguistics, or of artificial intelligence, and even core computer science. To the extent that the public thinks at all about what we do, they think of us as producing gadgets, such as amusing new sorts of telephone. Only the other day, a colleague was called up by someone in a neighboring department asking if we could mend his PC for him. ("Don't you have a Little Man? You *used* to have a Little Man.")

As a result, computer science is continually subjected to governmental reviews seeking assurance that we know what we are doing, and are doing enough for the economy. Many of these reviews draw very negative conclusions—the 1966 report of the Automatic Language Processing Advisory Committee (ALPAC) of the US National Academy of Sciences effectively shut down research in machine translation for over a decade, and was the main reason for ACL changing its name in 1968 from the Association for Machine Translation and Computational Linguistics (AMTCL). The 1973 report of Sir James Lighthill (a fluid dynamicist known for his foundational work in the field of aeroacoustics) to the UK Science Research Council (SRC) closed down artificial intelligence and NLP research for a decade, until the Alvey report decided that British industry had fallen behind in this area, and opened it up again. There have been many similar examples since then, though few as catastrophic.

Nobody goes around telling physicists what not to work on, or setting up commissions chaired by complete outsiders (roboticists or computational linguists, perhaps?) to decide whether physicists are earning their keep. The physicists tell the goverment what they think it is right to do, and the government either funds it or it doesn't. Even when it doesn't, as in the case of the superconducting supercollider, it's because they can't afford it, or lack the political power, not because of low esteem. How do the physicists do it?

Of course, as Duke Ellington said when asked how he kept his band together through the Beatles' era, you have to have a gimmick. The physicists gave us atomic energy and the bomb, so no one can ever suggest again that they do not deliver Bang for the Buck, *even when they actually don't*, as may well turn out to be the case for the past twenty years or so of research in string theory and supersymmetry.[1]

However, it is far more important that physics consists of a body of great empirically proven laws that all scientists recognize, from the laws of thermodynamics to the special and general theories of relativity and quantum theory. This body of knowledge lends both authority, and a breadth of vision that transcends any individual physicist's work and any individual theory, even if parts of it can be temporarily ignored when convenient.

But *we too* have awesome devices. Search engines have arguably changed people's lives at least as profoundly as atomic energy. The statistical machine translation tools that Google launched around May 2006 with Arabic for all the world to freely beta-test, and which have since been extended to Chinese, Russian, Japanese, and Korean, imperfect as they are, may well have an even bigger impact. Our colleagues in AI rejoice in beating international chess Grand Masters with Deep Blue, and boast of robots on Mars and autonomous vehicles charging around the Mojave desert. Computer science has the Internet itself to show off.

We too have discovered great truths—Zipf's Law, Information Theory, the power of statistically approximate language models, the only-just-trans-context-free automata-theoretic level of natural languages, the surface-compositionality of natural language

1 For a recent review of this question, see Smolin 2006.

semantics, and have dreams of a grand unified head dependency-driven Theory of Everything that that will one day make probabilistic and deterministic components work together to yield Meaning. How come *we* don't get respect?

The answer, I'm afraid, is that we haven't yet managed to form ourselves into a discipline, within which those important truths are respected and held beyond reasonable doubt, in whose name we stand united in the eyes of the world. Instead, our history has been sectarian, with dominant factions seeking to suppress proposals that diverge too much from the current orthodoxy, until they too are overthrown.

This tendency is very evident in computer science peer review of proposals to funding agencies, which are notably harsher than in other sciences, and which have the effect of making the agencies assume a shortage of good science in the field, so that they reduce funding accordingly. Worse still, when we have theoretical disagreements, we play them out in public. (The split between connectionists and the rest of AI/NLP is a case in point. It is sad to note that similar internal dissension seems to have played a role in the adverse conclusions reached in both the ALPAC and Lighthill reports.)

The physicists don't do this. When they review proposals, they rate them excellent at around twice the rate in CS, even if it means not getting funded themselves in the current round. The funding agencies conclude that there is a surplus of good science there, and seek or allocate further funding for the next round. When there are disagreements—as there seem to be currently about string theory—they are argued out behind closed doors, until a consensus can be reached and be presented to the general public. The biologists behaved similarly over a recent disagreement concerning methods for sequencing the genome.

One measure of our own fisiparous tendencies is that, while from the 1950s to the 1980s, the information theoreticians and statistical modelers among us used to make common cause with the linguists, we have subsequently drifted apart. We have come to believe that the linguists have forgotten Zipf's law, which says that most of the *variance in language behavior* can be captured by a small part of the system.

The linguists, on the other hand, think that it is we who have forgotten Zipf's law, which also says that most of the *information about the language system as a whole* is in the Long Tail.

It is we who are at fault here, because the machine learning techniques that we rely on are actually very bad at inducing systems for which the crucial information is in rare events—like physics, for example.

There is a grave danger here for the future of our putative discipline. One day, either because of the demise of Moore's law, or simply because we have done all the easy stuff, the Long Tail will come back to haunt us. For example, consider the current state of our former nemesis MT.

2. Machine Translation Then and Now

Machine translation (MT) was one of the earliest applications envisaged for computing machinery. Weaver (1949) identified the extreme ambiguity of natural language as the central problem of MT, and outlined as possible solutions not only the "noisy channel" model of SMT and its basis in *n*-gram language models, investigated with Claude Shannon (Shannon and Weaver 1949), but also the notion of parsing as logical deduction, and the interlingua-based syntax-driven approach to MT, which he based on the linguists' notion of Universal Grammar. (No sectarianism *there*, at least.)

It is interesting that, in that bright dawn, MT seemed likely to prove one of the easier cognitive tasks to mimic by machine. There seemed to be a real possibility that

there might be ways to bypass understanding altogether using simple purely syntactic and probabilistic devices. Over the next few years, there was a gradual disillusion with this expectation, for at least three reasons.

First, Chomsky (1957) showed beyond a reasonable doubt that natural languages not only fell outside the class that could be completely recognized using finite-state machines of the class implicit in Shannon's models, but also fell outside the class that could be adequately represented by context-free grammars. It wasn't immediately clear how to translate such grammars into computational terms, and they appeared to break the unity of grammar and probabilistic model that had been one of the attractions of the information-theoretic approach (although Chomsky was careful to acknowledge the possibility of statistically approximating such grammars with Markov processes for the purpose of reducing ambiguity).

Second, it became clear that the grammars that were needed for real texts, even within narrow genres like newspaper text or scientific papers, were very large indeed. The lack of computational transparency of grammar formalisms tended to give the hand-built grammars of this period the character of unstructured programs, with no apparent prospect of inducing them automatically.

Third, the problem of ambiguity for grammars of this size was much worse than anyone had expected.

As a result, opinion changed, and MT became seen as one of the hardest cognitive problems, requiring deep understanding of semantics and knowledge of the world, as well as full syntactic processing.

There is an often-repeated story about these days (one which has all the hallmarks of a myth) concerning a demonstration of an early Russian dictionary-based MT system using back-translation from English to Russian and back again. According to legend, the demonstrators were disconcerted to find that the sentence *Time flies like an arrow* returned as *Time flies enjoy arrows,* revealing a syntactic analysis of the sentence that would have been appropriate for the sentence *Fruit flies like a banana.*[2]

Of course we now know that the problem of the 1950s and 1960s was the lack of language models, of a kind that is now commonplace, thanks to Moore's law and HMMs, together with a method for learning such models and integrating them with structural rules. Some real progress has been made, and MT is once again thought of as an "easy" problem, that can be at least partly solved with relatively low-level methods.

3. An Experiment

So, how much progress have we made? We can repeat the back-translation experiment with Google Language Tools Beta *n*-gram-and-finite-state-transducer-based Arabic SMT. (The first line is the English input, the Arabic is its SMT translation, the third line is a gloss of the Arabic words, and the last line is the result of translating the Arabic back again by SMT.)

```
Time flies like an arrow.
الوقت الذباب يشبه السهم.
time-DEF flies-DEF resemble arrow-DEF
"Time flies like arrow."
```

2 There are many variants of this apocryphal story, involving other examples, all of which are likely to have their origin in much older jests about *human* dictionary-based translation (Hutchins 1995).

On the basis of the result of the back-translation, this looks OK, give or take a bit of morphology, but of course end-to-end back-translation is a very weak test, where you can just get lucky. Readers of Arabic will notice that the translation of *like* is indeed a comparative, not a verb meaning *enjoy*, as in the legend. However, they will also notice that *flies* translates as the noun, rather than the verb, just as the story foretold.

If we try the same test on *Fruit flies like a banana*, the flies are still insects, but *like* is still comparative, rather than a verb. So the two sentences are analyzed the same way, as in the story.

```
Fruit flies like a banana.
ذباب الفاكهة مثل الموز.
Flies-N fruit-N resemble banana-N
"Fruit flies like bananas."
```

Of course, all this is *very unfair*, and not at all surprising. The examples are out of domain, so the language model doesn't help us at all. So let's try an in-domain example of newswire text.

The following is almost the first text I found by searching Arabic Web pages for the Arabic for "Google Machine Translation," simply because I had already read the English reference document,[3] and I was pretty sure it would be out there somewhere. It is a human-authored Arabic translation of a recent Reuters story about the launching of Google Language Tools, taken from Al Jazeera:[4]

> ويقوم فرانز أوتش وهو ألماني يقود جهود غوغل الخاصة بالترجمة بتغذية الحاسوب بمئات الملايين من الكلمات من نصوص موازية مثل العربية والإنجليزية مستخدما وثائق الأمم المتحدة والاتحاد الأوروبي مصادر رئيس.
>
> وعن طريقة الترجمة الجديدة قال أوتش إنه رغم أن الجودة لن تكون كاملة يعد ذلك تطورا في المساعي السابقة الخاصة بالترجمة الآلية، وإن الترجمة الصحيحة في أغلبها قد تكون جيدة بما يكفي لبعض المهام. وذكر أنه كلما زادت البيانات التي يتم تغذية النظام بها كانت النتائج أفضل.
>
> وأثنى مايلز أوسبورن الأستاذ بجامعة أدنبره الذي قضى العام الماضي في العمل في مشروع غوغل على جهود الشركة، غير أنه لفت إلى أن البرمجيات لن تتغلب على البشر في الترجمات الماهرة كما تفعل في لعبة الشطرنج وأنه ينبغي استخدام البرمجيات للفهم وليس لإنجاز وثائق.

Here is the SMT translation, delivered in about the time it would take a native speaker to read the original:

> The German Franz Ouch which leads efforts Google translation computer feeds hundreds of millions of words of parallel texts such as Arabic,

3 http://www.reuters.com/article/technologyNews/idUSN1921881520070328?feedType=RSS%20target=.
4 http://www.aljazeera.net/news/archive/archive?ArchiveId=1037015.

> English, using documents of the United Nations and the European Union key sources.
>
> And how a new translation Ouch said that although the quality would not be complete That was a good in the previous translation mechanism, and that the correct translation mostly might be good enough for some tasks. He stated that more data be fed by the results were better.
>
> . . . He commended Miles Osborne Professor at the University of Edinburgh, who died last year at work in the company's efforts to Google, but he pointed out that the software will not prevail over people skilled in translations as they do in the game of chess and should use software to understand and not to complete documents.

The first thing to notice is that this is really very good. It is quite clear what the story is, and you can even guess that what Franz Och actually said in the English reference text was: "The more data we feed into the system, the better it gets." It even seems to know that "Google" can be a verb.

However, it exhibits all the problems to which we have always known MT is heir. Both pronouns "he" in the last paragraph will be understood as referring to Franz, whereas in the reference text it is Miles Osborne who does the commending and the pointing out. Moreover, the alarming rumor of the latter's death has been greatly exaggerated by the English language-model: The reference text says he "spent a sabbatical last year working on the Google project." The human Arabic translation says much the same, but the Arabic words for *spent* and *died* are homographs, and the newswire-based model favors the latter.

And of course, our friends the Construction Grammarians will gleefully point out that the system makes a hash of the unbounded dependency in Franz's use of what they call the "MORE-MORE" construction.

We can specifically probe the disability with respect to other kinds of unbounded dependencies, using back-translation on artificially generated (but in-domain) examples:

```
This is the bank that bought the company.
وهذا البنك هو ان اشترت الشركة.
"This is the bank that bought the company."

This is the company that the bank bought.
هذه هي الشركة التي اشترت البنك.
"*This is the company that bought the bank."

This is the bank that wants to buy the company.
هذا هو المصرف الذي يريد لشراء الشركة.
"This is the bank, which wants to buy the company."

This is the company which the bank wants to buy.
هذه هي الشركة التي تريد شراء البنك.
"*This is the company that wants to buy the bank."
```

Comparing the examples that are translated correctly and those (labeled *) that are not, the generalization is already clear: even a 5-gram model can only handle *root subject* relative clauses. Object relatives are beyond the horizon. (These effects are robust under variation of the content words.) Here are some more challenging embedded examples that confirm the diagnosis:

```
This is the company that said the bank bought bonds.
هذه هي الشركة التي قال البنك بشراء السندات.
"This is the company that said the bank bought the bonds."

This is the company that the bank said bought bonds.
هذه هي الشركة التي قال البنك بشراء السندات.
"*This is the company that said the bank bought the bonds."

These are the bonds that the company said that the bank bought.
هذه هي سندات الشركة أن البنك اشترى.
"*These are the bonds that the bank bought the company."
```

4. Who Cares?

What does this tell us? Nothing that we shouldn't have already known. We knew that *n*-gram models and FSTs weren't going to handle long-range dependencies, because Chomsky told us so. That's one of the Big Truths of computational linguistics.

Our situation is in fact rather like that of the physicists. We have one theory for talking about phenomena on a large scale, just as they have the General Theory of Relativity, and another theory for talking about the very small scale, just as they have Quantum Theory. Like the physicists, we have difficulty in reconciling those theoretical levels. Like them, some of us think it's fine to have two theories, whereas others of us think it's intolerable.

The former kind of computational linguist will point out that long-range dependencies are sparse. (There are around 20K `*T*` empty categories in around 16K of the roughly 40K sentences in the Penn Treebank, of which around 6K seem to be non-subject, non-sentential *wh*-traces of some kind.) Worrying about them isn't going to significantly impact overall parser dependency recovery rates, much less *n*-gram-precision-based BLEU scores. By the time we have fed enough data into the system to make it know that spending a sabbatical at Google is more likely than dying there, and Moore's law has made the machines exponentially bigger and faster, and fancier algorithms allow us to deal with bigger *n*-grams, maybe this problem will go away.

Maybe. These are certainly good reasons for the IR and SMT researchers to keep working at the large scale, giving the world these amazing search engines and translation aids that give human beings vastly increased access to other cultures. This is our discipline's equivalent in street-credibility terms of delivering atomic energy and robots on Mars.

Nevertheless, to the other kind of computational linguist, it sounds depressingly like getting better and better at recalling what is already well-known, and understanding what has often been said before.

They point out that, in the long run, finite state methods alone may simply not work. Accuracy in most areas (WER in ASR, BLEU score in SMT, Eval-b for parsers) is *at best*

linear in the logarithm of the amount of training data. Even optimistic extrapolation of current learning curves suggests truly incredible amounts of data will be needed (Lamel, Gauvain, and Adda 2002; Moore 2003; Knight and Koehn 2004; Brants et al. 2007).

Moreover, the more of the local stuff we get right, the more users will come to trust the software, and hence the more noticeable long range dependencies will become, and the more upset people will get if they are deceived by a wrong analysis.

What this should tell us is that the Long Tail is not mocked. Long-range dependencies of the kind investigated above are semantically crucial. Ignoring them disrupts all the *other* dependencies in those examples. (They are also more frequent in genres like questions.) So we need to remember—and above all, teach our students—what our discipline tells us the problem is, even when it's not doing much for our BLEU score. In this connection, it is encouraging to see that many of the MT papers in the 2007 ACL explicitly invoked syntax-level representations.

But we need to do more than this. To get respect, and avoid the risk of yet another AI/NLP Winter, we will need to pull ourselves together as a discipline, lift our eyes above our rivalries, and take a longer view of where we are going (and where we have come from) than seems to be the current style. This will probably require a gradual move to a more considered and authoritative style of publication, with journal articles taking the place of hastily written and reviewed conference papers, as another author of this column recently suggested.

It will mean speaking with one voice, as the physicists and biologists do, and supporting a diversity of views that transcends fashion and funding, wherever the science is good. It also means telling the public in honest terms how to think about what we do, what can be relied on and what the really hard problems are, in good times and bad. This should not be too difficult if we keep reminding them and ourselves of the following:

> Human knowledge is expressed in language. *So computational linguistics is very important.*

References

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large language models in machine translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 858–867, Prague.

Chomsky, Noam. 1957. *Syntactic Structures*. Mouton, The Hague.

Hutchins, John. 1995. "The whisky was invisible" or persistent myths of MT. *MT News International*, 11, 17–18.

Knight, Kevin and Philipp Koehn. 2004. What's new in statistical machine translation. Tutorial, HLT/NAACL 2004.

Lamel, Lori, J. L. Gauvain, and G. Adda. 2002. Unsupervised acoustic model training. In *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, pages 877–880, Orlando, FL.

Moore, Roger. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of Eurospeech Conference*, pages 2582–2585, Geneva.

Shannon, Claude and Warren Weaver. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.

Smolin, Lee, ed. 2006. *The Trouble with Physics*. Houghton Mifflin, New York.

Weaver, Warren. 1949. Translation. Manuscript, Rockefeller Institute. Reproduced as Weaver (1955).

Weaver, Warren. 1955. Translation. In William Locke and Andrew Booth, editors, *Machine Translation of Languages: Fourteen Essays*. MIT Press, Cambridge, MA, pages 15–23.