[From: *Current issues in computational linguistics: in honour of Don Walker*, ed. Antonio Zampolli, Nicoletta Calzolari, Martha Palmer (Linguistica Computazionale, vol. 9-10); Pisa, Dordrecht, [1994]

Stone Soup and the French Room

Yorick Wilks
Department of Computer Science
University of Sheffield
e-mail: yorick@dcs.sheffield.ac.uk

Abstract

The paper argues that the IBM statistical approach to machine translation has done rather better after a few years than many sceptics believed it could. However, it is neither as novel as its proponents suggest nor is it making claims as clear and simple as they would have us believe. The performance of the purely statistical system (and we discuss what that phrase could mean) has not equaled the performance of SYSTRAN. More importantly, the system is now being shifted to a hybrid that incorporates much of the linguistic information that it was initially claimed by IBM would not be needed for MT. Hence, one might infer that its own proponents do not believe “pure” statistics sufficient for MT of a usable quality. In addition to real limits on the statistical method, there are also strong economic limits imposed by their methodology of data gathering. However, the paper concludes that the IBM group have done the field a great service in pushing these methods far further than before, and by reminding everyone of the virtues of empiricism in the field and the need for large scale gathering of data.

1 History

Like connectionism, statistically-based machine translation is a theory one was brought up to believe had been firmly locked away in the attic, but here it is back in the living room. Unlike connectionism, it carries no psychological baggage, in that it seeks to explain nothing and cannot be attacked on grounds of its small scale as connectionist work has been. On the contrary that is how it attacks the rest of us.

> It is well known that Western Languages are 50% redundant. Experiment shows that if an average person guesses the successive words in a completely unknown sentence he has to be told only half of them. Experiment shows that this also applies to guessing the successive word-ideas in a foreign language. How can this fact be used in machine translation? (King, 1956).

Alas, that early article told us little by way of an answer and contained virtually no experiments or empirical work. Like IBM’s approach it was essentially a continuation of the idea underlying Weaver’s original memorandum on MT: that foreign languages were a code to be cracked. I display the quotation as a curiosity, to show that the idea itself is not new and was well known to those who laid the foundations of modern representational linguistics and AI.

I personally never believed Chomsky's arguments in 1957 against other theories than his own any more than I did what he was for: his attacks on statistical and behaviorist methods (as on every thing else, like phrase structure grammars) were always in terms of their failure to give explanations, and I will make no use of such arguments here, noting as I say that how much I resent IBM's use of "linguist" to describe everyone and anyone they are against. There is a great difference between linguistic theory in Chomsky's sense, as motivated entirely by the need to explain, and theories, whether linguistic/AI or whatever, as the basis of procedural, application-engineering-orientated accounts of language. The latter stress testability, procedures, coverage, recovery from error, non-standard language, metaphor, textual context, and the interface to general knowledge structures.

Like many in NLP and AI, I was brought up to oppose linguistic methods on exactly the grounds IBM do: their practitioners were uninterested in performance and success at MT in particular. Indeed, the IBM work to be described here has something in common with Chomsky's views, which formed the post-1957 definition of "linguist". It is clear from Chomsky's description of statistical and Skinnerian methods that he was not at all opposed to relevance/pragmatics/semantics-free methods – he advocated them in fact – it was only that, for Chomsky, the statistical methods advocated at the time were too simple a method to do what he wanted to do with transformational grammars. More recent developments in finite state (as in phrase structure) grammars have shown that Chomsky was simply wrong about the empirical coverage of simple mechanisms.

In the same vein he dismissed statistical theories of language on the ground that sentence pairs like:

the.
I saw a
triangular whole.

are equally unlikely but utterly different in that only the first is ungrammatical. It will be clear that the IBM approach discussed here is not in the least attacked by such an observation.

Is the debate about empiricism? No.

Anyone working in MT, by whatever method, must care about success, in so far as that is what defines the task. Given that, the published basis of the debate between rationalism and empiricism in MT is silly: we are all empiricists and, to a similar degree, we are all rationalists, in that we prefer certain methodologies to others and will lapse back to others only when our empiricism forces us to. That applies to both sides in this debate, a point I shall return to.

An important note before continuing: when I refer to IBM machine translation I mean only the systems referred to at the end by Brown et al. IBM as a whole supports many approaches to MT, including McCord's (1989) prolog-based symbolic approach, as well as symbolic systems in Germany and Japan.

Is the debate about how we evaluate MT? No.

In the same vein, I shall not, as some colleagues on my side of the argument would like, jump ship on standard evaluation techniques for MT and claim that only very special and sensitive techniques (usually machine-aided techniques to assist the translator) should in future be used to assess our approach.

MT evaluation is, for all its faults, probably in better shape than MT itself, and we should not change the referee when we happen not to like how part of the game is going. Machine-aided translation (MAT) may be fine stuff, but IBM's approach should be competed with head on by those who disagree with it. In any case, IBM's method could in principle provide, just as any other system could, the first draft translation for a translator to improve on line. The only argument against that is that IBM's would be a less useful first draft *if a user wanted to see why certain translation decisions had been taken*. It is a moot point how important that feature is. However, and this is a point Slocum among others has made many times, the evaluation of MT must in the end be economic not scientific. It is a technology and must give added value to a human task. The ALPAC report, it is often forgotten, was about the economics of contemporary MT, not about its scientific status: the report simply said that MT at that time was not competitive, quality for quality, with human translation.

SYSTRAN won that argument later by showing there was a market for the quality it produced at a given cost. We shall return to this point later, but I make it now because it is one that does tell, in the long run, on the side of those who want to emphasize MAT. But for now, and for any coming showdown between statistically and non-statistically based MT – where the latter will probably have to accept SYSTRAN as their champion for the moment, like it or not – we might as well accept existing "quasi-scientific" evaluation criteria, Cloze tests, test sets of sentences, improvement and acceptability judged by monolingual and bilingual judges, etc. None of us in this debate and this research community are competent to settle the economic battle of the future, decisive though it may be.

2 Arguments Not to Use Against IBM

There are other well known arguments that should not be used against IBM, such as that much natural language is mostly metaphorical and that applies to MT as much as any other NLP task and statistical methods cannot handle it. This is a weak but interesting argument: the awful fact is that IBM cannot even consider a category such as metaphorical use. Everything comes out in the wash, as it were, and it either translates or it does not and you cannot ask why. Much of their success rate of sentences translated acceptably is probably of metaphorical uses. There may be some residual use for this argument concerned with very low frequency types of deviance, as there is for very low frequency words themselves, but no one has yet stated this clearly or shown how their symbolic theory in fact gets such uses right (though many of us have theories of that). IBM resolutely deny the need of any such special theory, for *scale* is all that counts for them.

3 What is the State of Play Right Now?

Away with rumor and speculation; what is the true *state of play* at the moment? In recent reported but unpublished DARPA-supervised tests the IBM system CANDIDE did well, but significantly worse than SYSTRAN's French-English system over texts on which neither IBM nor SYSTRAN had trained. Moreover, CANDIDE had far higher standard deviations than SYSTRAN, which is to say that SYSTRAN was far more consistent in its quality (just as the control human translators had the lowest standard deviations across differing texts). French-English is not one of SYSTRAN's best systems but this is still a significant result. It may be unpleasant for those in the symbolic camp, who are sure their own system could, or should, do better than SYSTRAN, to have to cling to it in this competition as the flagship of symbolic MT, but there it is. IBM have taken about 4 years to get to this point. French-English SYSTRAN was getting to about IBM's current levels after 3-4 years of work. IBM would reply that that they are an MT system factory, and could do the next language much faster. We shall return to this point.

4 What is the Distinctive Claim by IBM About How to Do MT?

We need to establish a ground zero on what the IBM system is: their rhetorical claim is (or perhaps was) that they are a pure statistical system, different from their competitors, glorying in the fact that they did not even need French speakers. By analogy with Searle's Chinese Room, one could call theirs a French Room position: MT without a glimmering of understanding or even knowing that French was the language they were working on! There is no space here for a detailed description of IBM's claims (see Brown et al., 1990, 1991a, 1991b). In essence, the method is an adaptation of one that worked well for speech decoding (Jelinek and Mercer, 1980).

The method establishes three components: (a) a trigram model of English sequences; (b) the same for French; (c) a model of quantitative correspondence of the parts of aligned sentences between French and English. The first two are established from very large monolingual corpora in the two languages, of the order of 100 million words, the third from a corpus of *aligned* sentences in a parallel French-English corpus that are translations of each other. All three were provided by a large machine-readable subset of the French-English parallel corpus of Canadian parliamentary proceedings (Hansard). (1) and (2) are valuable independent of the language pair and could be used in other pairings, which is why they now call the model a *transfer* one. A very rough simplification: an English sentence yields likeliest equivalences for word strings (sub-strings of the English input sentence), i.e., French word strings. The trigram model for French re-arranges these into the most likely order, which is the output French sentence. One of their most striking demonstrations is that their trigram model for French (or English) reliably produces (as the likeliest order for the components) the correct ordering of items for a sentence of ten words or less.

What should be emphasized is the enormous amount of pre-computation that this method requires and, even then, a ten word sentence as input requires an additional hour of computation to produce a translation. This figure will undoubtedly reduce with time and hardware expansion but it gives some idea of the computational intensity of IBM's method.

The facts are now quite different. They have taken in whatever linguistics has helped: morphology tables, sense tagging (which is directional and dependent on the properties of French in particular), a transfer architecture with an intermediate representation, plural listings, and an actual or proposed use of bilingual dictionaries. In one sense, the symbolic case has won: they topped out by pure statistics at around 40% of sentences acceptably translated and then added whatever was necessary from a symbolic approach to upgrade the figures. No one can blame them: it is simply that they have no firm position beyond taking what ever will succeed, and who can object to that?

There is then no theoretical debate at all, and their rhetorical points against symbolic MT are in bad faith. It is Stone Soup: the statistics are in the bottom of the pot but all flavor and progress now come from the odd trimmings of our systems that they pop into the pot.

They are, as it were, wholly pragmatic statisticians: less pure than, say, the Gale group (e.g., Gale & Church 1990) at AT&T: this is easily seen by the IBM introduction of notions like the one they call “informants” where a noun phrase of some sort is sought before a particular text item of interest. This is an interpolation of a highly theoretically-loaded notion into a routine that, until then, had treated all text items as mere uninterpreted symbols.

One could make an analogy here with localist versus distributivist sub-symbolic connectionists: the former, but not the latter, will take on all kinds of categories and representations developed by others for their purposes, without feeling any strong need to discuss their status as artifacts, i.e., how they could have been constructed other than by handcrafting.

This also makes it hard to argue with them. So, also, does their unacademic habit of telling you what they've done but not publishing it, allegedly because they are (a) advancing so fast, and (b) have suffered ripoffs. One can sympathize with all this but it makes serious debate very hard.

5 The Only Issue

There is only one real issue: is there any natural ceiling of success to *PURE* statistical methods? The shift in their position suggests there is. One might expect some success with those methods on several grounds (and therefore not be as surprised as many are at their success):

- There have been substantial technical advances in statistical methods since King’s day and, of course, in fast hardware to execute such functions, and in disk size to store the corpora.

- The redundancy levels of natural languages like English are around 50% over both words and letters. One might expect well-optimized statistical functions to exploit that to about that limit, with translation as much as another NLP task. One could turn this round in a question to the IBM group: how do they explain why they get, say, 40-50% or so of sentences right, rather than 100%? If their answer refers to the well-known redundancy figure above, then the ceiling comes into view immediately.

If, on the other hand, their answer is that they cannot explain anything, or there is no explaining to do or discussions to have, then their task and methodology is a very odd one indeed. Debate and explanation are made impossible and, where that is so, one is normally outside any rational or scientific realm. It is the world of the witch-doctor: Look – I do what I do and notice that (sometimes) it works.

- According to a conjecture I propounded some years ago, with much anecdotal support, *any theory whatever no matter how bizarre will do some MT.* Hence my surprise level is always low.

6 Other Reasons for Expecting a Ceiling to Success with Statistics

Other considerations that suggest there is a ceiling to pure statistical methods are as follows:

1. A parallel with statistical information retrieval may be suggestive here: it generally works below the 80% threshold, and the precision/recall tradeoff seems a barrier to greater success by those methods. Yet it is, by general agreement, an easier task than MT and has been systematically worked on for over 35 years, unlike statistical MT whose career has been intermittent. The relationship of MT to IR is rather like that of sentence parsers to sentence recognizers. A key point to note is how rapid the early successes of IR were, and how slow the optimization of those techniques has been since then!

2. A technical issue here is the degree of their reliance on alignment algorithms as a pre-process: in ACL91 they claimed only 80% correct alignments, in which case how could they exceed the ceiling that that suggests?

3. Their model of a single language is a trigram model because moving up to even one item longer (i.e., a quadgram model) would be computationally prohibitive. This alone must impose a strong constraint on how well they can do in the end, since any language has phenomena that connect outside the three item window. This is agreed by all parties. The only issue is how far one can get with the simple trigram-model (and, as we have seen, it gives a basic 40%), and how far can distance phenomena in syntax be finessed by forms of information caching. One can see the effort to extend the window as enormously ingenious, or patching up what is a basically inadequate model when taken alone.

7 The Future: Hybrid Approaches

Given the early success of IBM's methods, the most serious and positive question should be what kinds of *hybrid* approach will do best in the future: coming from the symbolic end, plus statistics, or from a statistical base but inducing, or just taking over, whatever symbolic structures help? For this we can only watch and wait, and possibly help a little here and there. However, there are still some subsidiary considerations.

7.1 IBM, SYSTRAN, and the Economics of Corpora

In one sense, what IBM have done is partially automate the SYSTRAN construction process: replacing laborious error feedback with statistical surveys and lexicon construction. And all of us, including SYSTRAN itself, could do the same. However, we must always remember how totally tied IBM are to their Hansard text, the Rosetta Stone, one might say, of modern MT. We should remember, too, that their notion of word sense is only and exactly that of correspondences between different languages, a wholly unintuitive one for many people.

The problem IBM have is that few such vast bilingual corpora are available in languages for which MT is needed. If, however, they had to be constructed by hand, then the economics of what IBM has done would change radically. By bad luck, the languages for which such corpora are available are also languages in which SYSTRAN already has done pretty well, so IBM will have to overtake, then widen the gap with, SYSTRAN's performance a bit before they can be taken seriously from an economic point of view. They may be clever enough to make do with less than the current 100 million word corpora per language, but one would naturally expect quality to decline as they did so.

This resource argument could be very important: Leech has always made the point, with his own statistical tagger, that any move to make higher-level structures available to the tagger always ended up requiring much more text than he had expected.

This observation does not accord with IBM's claims, which are rather the reverse, so an important point to watch in future will be whether IBM will be able to obtain adequate bilingual-corpora for the domain-specialized MT that is most in demand (such as airline reservations or bank billings). Hansard has the advantage of being large but is very very general indeed.

7.2 Why the AI Argument About MT Still Has force

The basic AI argument for knowledge-based processing does not admit defeat and retreat, it just regroups. It has to accept Bar Hillel's old anti-MT argument (Bar Hillel, 1960) on its own side – i.e., that as he said, good MT must in the end need knowledge representations. One version of this argument is the primitive psychological one: humans do not do translation by exposure to such vast texts, because they simply have not had such exposure, and in the end how people do things will prove important. Note that this argument makes an empirical claim about human exposure to text that might be hard to substantiate. This argument will cut little ice with our opponents, but there may still be a good argument that we do need representations for tasks in NLP related to MT: e.g. we cannot really imagine doing summarization or question answering by purely statistical methods, can we? There is related practical evidence from message extraction: in the MUC competitions (Lehnert & Sundheim, 1991), the systems that have done best have been hybrids of preference and statistics, such as of Grishman and Lehnert, and not pure systems of either type.

There is the related argument that we need access to representations *at some point* to repair errors. This is hard to make precise but fixing errors makes no sense in the pure IBM paradigm; you just provide more data. One does not have to be a hard line syntactician to have a sense that rules do exist in some linguistic areas and can need fixing.

7.3 Hard Problems Do Not Go Away

There remain, too, crucial classes of cases that seem to need symbolic inference: an old, self-serving, one will do such as “The soldiers fired at the women and I saw several fall” (Wilks, 1975).

I simply cannot imagine how any serious statistical method (e.g., not like “pronouns are usually male so make “several” in a gendered translation agree with soldiers”!) can get the translation of “several” into a gendered language right (where we assume it must be the women who fall from general causality). But again, one must beware here, since presumably any phenomenon whatever will have statistically significant appearances and can be covered by some such function if the scale of the corpus is sufficiently large. This is a truism and goes as much for logical relations between sentences as for morphology. It does not follow that that truism leads to tractable statistics or data gathering. If there could be 75,000-word-long Markov chains, and not merely trigrams (which seem the realistic computational limit) the generation of whole novels would be trivial. It is just not practical to have greater-than-three chains but we need to fight the point in principle as well!

Or, consider the following example (due to Sergei Nirenburg):

> PRIEST IS CHARGED WITH POPE ATTACK
> (Lisbon, May 14)
>
> A Spanish priest was charged here today with attempting to murder the Pope. *Juan Fernandez Krohn*, aged 32, was arrested after *a man armed with a bayonet* approached the Pope while he was saying prayers at Fatima on Wednesday night.
>
> According to the police, *Fernandez* told the investigators today he trained for the past six months for the assault. He was alleged to have claimed the Pope ‘looked furious’ on hearing *the priest’s* criticism of his handling of the church’s affairs. If found guilty, *the Spaniard* faces a prison sentence of 15-20 years.
>
> (*The Times* 15 May 1982)

The five italicized phrases all refer to the same man, a vital fact for a translator to know since some of those phrases could not be used in any literal manner in another language (e.g. “the Spaniard” could not be translated word-for-word into Spanish or Russian). It is hard to imagine multiple identity of reference like that having *any* determinable statistical basis.

8 Is the Pure Statistics Argument What is Being Debated? No

Everything so far refers to the *pure statistics argument*, from which IBM have now effectively backed off. If the argument is then to be about the deployment of hybrid systems and exactly what data to get from the further induction of rules and categories with statistical functions (e.g., what sort of dictionary to use) then there is really no serious argument at all, just a number of ongoing efforts with slightly differing recipes. Less fun, but maybe more progress, and IBM are to be thanked for helping that shift.

8.1 IBM as Pioneers of Data Acquisition

I can add a personal note there: when I worked on what I then called Preference Semantics (Wilks, 1975) at McCarthy's Stanford AI Lab, McCarthy always dealt briefly with any attempt to introduce numerical methods into AI – statistical pattern-matching in machine vision was a constant irritation to him – by saying "Where do all those numbers COME from?" I felt a little guilty as Preference Semantics also required at least link counting. One could now say that IBM's revival of statistical methods has told us exactly where some of these numbers come from! But that certainly does not imply that the rules that express the numbers are therefore useless or superseded.

This touches on a deep metaphysical point: I mentioned above that we may feel word-sense is a non-bilingual matter, and that we feel that there *are* rules that need fixing sometimes, and so on. Clearly, not everyone feels this. But it is our culture of language study that tells us that rules, senses, metaphors, representations etc. are important and that we cannot imagine all that is just a cultural artifact. An analogy here would be Dennett's recently (1991) restated theory of human consciousness that suggests that all our explanations of our actions, reason, motives, desires etc. as we articulate them may be no more than fluff on the underlying mechanisms that drive us.

IBM's work induces the same terror in language theorists, AI researchers and linguists alike: all their dearly-held structures may be just fluff, a thing of schoolmen having no contact with the reality of language. Some of us in AI, long ago, had no such trouble imagining most linguistics was fluff, but do not want the same argument turned round on us, that *all* symbolic structures may have the same status.

Another way of looking at this is how much good IBM are doing us all: by showing us, among other things, that we have not spent enough time thinking about how to acquire, in as automatic a manner as possible, the lexicons and rule bases we use. This has been changing lately, even without IBM's influence, as can be seen from the large-scale lexical extraction movement of recent years. But IBM's current attempts to recapitulate, as it were, in the ontogeny of their system, much of the phylogeny of the AI species is a real criticism of how some of us have spent the last twenty years.

We have not given enough attention to knowledge acquisition, and now they are doing it for us. I used to argue that AIers and computational linguists should not be seen as the white-coated laboratory assistants of linguistic theorists (as some linguists used to dream of using us). Similarly, we cannot wait for IBMers to do this dirty work for us while we go on theorizing. Their efforts should change how the rest of us proceed from now on.

9 Conclusion: Let Us Declare Victory and Carry on Working

Relax, go on taking the medicine. Brown et al.'s retreat to incorporating symbolic structures show the pure statistics hypothesis has failed. All we should be haggling about now is how best to derive the symbolic structures we use, and will go on using, for machine translation.

Acknowledgments
In acknowledgement of contributions from James Pustejovsky, Bob Ingria, Bran Boguraev, Sergei Nirenburg, Ted Dunning and others in the CRL natural language processing group.

References

[1] Bar-Hillel, Y., "The present status of automatic translation of languages", in J. Alt (ed.), *Advances in Computers* 1, Academic Press, New York, 1960.

[2] Brown, P.F., J. Cocke, S. DellaPietra, V. DellaPietra, F. Jelinek, J. Lafferty, R. Mercer, P. Roossin, "A statistical approach to machine translation", in *Computational Linguistics*, 16, 1990,79-85.

[3] Brown, PR, J. Lai, R. Mercer, "Aligning sentences in parallel corpora", in *Proceedings 29th Annual Meeting of the Association for Computational Linguistics,* Berkeley, CA, 1991, 169-176.

[4] Chomsky, N., *Syntactic Structures,* Mouton and Co., The Hague, 1957.

[5] Dennett, D., *Consciousness Explained,* Bradford Books, Cambridge MA, 1991.

[6] Gale, W., K. Church, "Poor estimates of context are worse than none", in *Proc. 1990 DARPA Speech and Language Meeting,* Hidden Valley, PA, 1990.

[7] King, G. "Stochastic methods of mechanical translation", in *Mechanical Translation,* 3, 1956.

[8] Jelinek, F., R. Mercer, "Interpolated estimation of Markov source parameters from sparse data", in *Proceedings of the Workshop on Pattern Recognition in Practice,* North Holland, Amsterdam, The Netherlands, 1980.

[9] Lehnert, W., B. Sundheim, "A performance evaluation of text analysis technologies", *AI magazine,* 12, 1991.

[10] McCord, M., "A new version of the machine translation system LMT", *Literary & Linguistic Computing,* 4, 1989.

[11] Wilks, Y, "A preferential pattern-matching semantics for natural language understanding", *Artificial Intelligence,* 11, 1975.