The Trend towards Statistical Models in Natural Language Processing

Mark Y. Liberman

Department of Computer and Information Science, University of Pennsylvania

1 A Flowering of Corpus-Based Research

Over the past few years, we have seen a significant increase in the number and sophistication of computational studies of large bodies of text and speech. Such studies have a wide variety of topics and motives, from lexicography and studies of language change, to methods for automated indexing and information retrieval, tagging and parsing algorithms, techniques for generating idiomatic text, cognitive models of language acquisition, and statistical models for application in speech recognizers, text or speech compression schemes, optical character readers, machine translation systems, and spelling correctors.

1.1 Aims and Applications

Although in some cases the corpus serves only as a source of heuristic examples or of test materials for evaluation, more often the result of such studies is a statistical model of some aspect of language, which can then be used as a tool for a variety of purposes. Typical applications include decoding messages in noise (speech recognition, optical character recognition, etc.), resolution of inherent analysis ambiguities (lexical category ambiguities, constituent structure ambiguities, ambiguities of sense and reference), similarity measures among chunks of text (information retrieval, message routing), low bit rate coding, and derivation of various sorts of lexicons. As these examples suggest, engineering applications have been in the lead, with the current interest of scientists still marginal, although growing, especially among researchers interested in language change and language learning, and among those who study resolution of ambiguities in human speech and language processing. Thus in this area, the IEEE has been ahead of the ACL, which in turn has been ahead of the LSA.

1.2 An Example of Statistical Modeling in Linguistic Processing

Many of the applications in pattern recognition can be viewed as specifying (implicitly or explicitly) a set of "theories" $\{T_i\}$, one of which will be invoked

to explain some particular observational evidence E_j. Then the recognition task becomes to find the theory T_i whose conditional probability given the evidence E_j is greatest. This is typically done via Bayes' Rule, setting

$$P(T_i|E_j) = \frac{P(E_j|T_i)P(T_i)}{P(E_j)}, \tag{1}$$

on the basis that the values on the right side of the equation are usually easier to estimate than the crucial quantity on the left.

For instance, in a simple model used to correct typing errors, the "theories" would be possible strings of "true," originally-intended letters; the "evidence" would be the string of letters actually typed; the term $P(E_j|T_i)$ represents a statistical model of the generation of errors in the process of typing (this is sometimes called the "channel" model, reflecting early applications in communications theory); and the term $P(T_i)$ is an estimate of the a priori probability of a hypothetical "true" letter string. This last term might reflect arbitrarily complex expectations about the material being typed, including its linguistic structure, its topic, and so forth; the function used to estimate this quantity is often called a "language model," or (again from usage in communications theory) a "source model." In this application, as in many others, the term $P(E_j)$, may be ignored, since it is the same for all theories.

A more complex instance of essentially the same structure is involved in most contemporary speech-recognition systems, with the "evidence" being a sequence of classes of noises rather than a typed string, and the "channel" being a model of the process of speaking rather than a model of typing-error generation. The "language model" might well be nearly the same in both cases, although we might also decide to exploit the significant differences between the two sorts of language at issue.

In either case, the expression "language model" is a bit misleading, since we are estimating the overall probability of a typed or spoken phrase, which depends heavily on issues that at best partly linguistic. For instance, both *last* and *lost* are adjectives, and thus could modify the noun *year*, but *last year* occurs in news-wire text more than 300 times per million words, while *lost year*, although perfectly well-formed and even sensible, is vanishingly unlikely. What is usually *lost* is *ground*, *souls*, *productivity*, or *wages*, while *ground*, if not *lost*, is likely to be *high*. Such collocational regularities are a mixture of facts about words and facts about concepts, topics and styles.

2 Some Historical Observations

There was a previous flowering of work on statistical models of natural language, and linguistic inference from corpora, in the 1950s and early 1960s. During the 1970s and early 1980s, the level of attention declined, especially among scientists but also to a considerable extent among engineers.

Thus Miller and Chomsky's monograph *Finitary Models of Language Users*, which appeared in 1963, had 43 pages on "stochastic models" versus 19 pages on "algebraic models," demonstating the importance that stochastic models had for

scientists as well as engineers up to that time. By contrast, Osherson, Strob and Weinstein's important book *Systems that Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists*, published in 1986, has 8 (of 205) pages devoted to what is called "A topological perspective," which (a bit shyly) sketches some of the issues that arise in learning languages on which a measure (such as a probability function) is defined. The word *stochastic* is not in this book's index. Even more strikingly, Partee, ter Meulen and Wall's monumental *Mathematical Methods in Linguistics*, published in 1990, has only one mention of statistical issues in its 663 pages, namely the point in the introduction where they observe that "we have not tried to cover probability..."

It's easy to offer explanations for this development. The early stochastic models (and indeed their modern counterparts) are sometimes breathtakingly naive. Often this represents a conscious (and proper) decision to see what can be done with a maximally simple, if obviously wrong, set of assumptions, but such a move can easily be misunderstood and subjected to ridicule by members of a rival technical culture. Many stochastic models of natural language can also be faulted on logical grounds, for not distinguishing among the conceptually different sorts of information contained in syntactic, semantic and pragmatic constraints. In *Syntactic Structures*, Chomsky presented an effective critique, along these lines, of the whole enterprise of frequentistic analysis of natural language, in which connection the famous *colorless green ideas* examples arose.

Until recently, in any case, even the simplest stochastic models were not economically practical for everyday applications, because of the high cost of the computer resources required to develop them and to use them. Furthermore, starting in the late 1950s, there was a lot of work for both scientists and engineers to do in exploring the higher levels of the just-discovered Chomsky hierarchy, and in trying to create and integrate models of linguistic meaning, world knowledge, and common-sense reasoning. At a more general level, we might also point to an anti-empiricist, anti-numerical, pro-symbolic trend in the *Zeitgeist* during those years. Counting things was just not seen as proper work for a gentleperson.

For all these reasons, interest in stochastic models and in corpus-based linguistic inference declined drastically. But meanwhile, in the scattered cells of what John Bridle has called the "Cybernetic Underground," engineers were developing practical applications that incorporated statistical models of natural language. The microelectronic revolution has made such applications genuinely practical; at the same time, it has become increasingly clear that research on knowledge-based approaches to speech and natural-language processing will not by itself produce effective broad-coverage programs. For these reasons, by the mid 1980s the field was ready to try another round of frequentistic research.

Speech recognition research led the way – the guerrillas of the cybernetic underground, such as Jim Baker and Fred Jelinek, had established their base camps in this area – and the efforts of DARPA to impose quantitative evaluation measures on its contractors played a crucial role. Probability theory, after all, originally arose to tell us how to "play the odds" when making decisions under circumstances of uncertainty; and whether in shooting craps or in recognizing speech, gamblers who know the odds, and place their bets accordingly,

will generally beat those who don't.

2.1 What's Really New?

As we have just suggested, the main motive force in the resurgence of corpus-based research has been the falling cost of computer technology, which makes complex speech and natural language systems affordable, and the fact that speech and natural language systems, which must resolve many ambiguities, perform much more accurately if they make their choices based on empirically-estimated odds.

In addition, there are a few new mathematical techniques that were not known during the 1950s, such as the re-estimation methods for Hidden Markov Models and Stochastic Context-Free Grammars, some techniques for inducing stepwise-optimal decision trees, and improved estimation procedures for dealing with the sparse data characteristic of linguistic distributions.

At least some of the models, this time around, are much more sophisticated: many of the the insights from thirty years of research on algebraic models of natural language are being adapted and used, either explicitly or as part of a common, default perspective on the problems. Where the statistical models are over-simplified, the false assumptions at issue are now more likely to be explicitly justified. It remains to be shown that statistical models with a more realistic architecture can be made to pay off, but the effort to find out is certainly under way in many laboratories.

Finally, everything is being done on a much larger scale. Claude Shannon made his 1951 estimates of the entropy of English text based on guesses at a few hundred letters; and in the 1960s, a million words was a large corpus; whereas Brown et al. (1990) base an estimate of the entropy of English text on the cross-entropy of a model based on almost 380 million words with an independent test corpus of a million words.

3 Why Corpora?

We can motivate the use of statistical models in speech and natural language processing simply by a desire to make optimal guesses when we don't know the answer for sure. But do we really need such enormous corpora, as opposed to (say) cleverer extrapolation from smaller bodies of evidence by means of better theories?

Certainly no one would argue against better theories, which are needed without any question, but there also seem to be good arguments for more data.

For one thing, purely as a practical matter, today's theories work better with more data, and so acquiring more data is a reliable and safe way to improve performance. One reason for this is probably that (as noted earlier) we are to some extent using collocational regularities to model regularities of the world rather than of speech and language; our models are learning about the world through talk and reading rather than through direct experience. From an

engineering point of view, this is a good thing, since we do not have any other reliable current prospects for approximating in broad domains the effects of world knowledge, real-world experience, and common-sense reasoning.

In any case, human linguistic experience is at least as large as the corpora that we are starting to work with now. A simple calculation suggests that people ordinarily hear at least 20 million spoken words a year; and a literate person whose job involves producing and interpreting text may easily read another 20 million written words. All of this suggests that a hundred million words is a reasonable size for a corpus of speech or text intended to model the linguistic experience of a linguistically-adept human.

4 Conclusion

On one view, effective models of human language use need not contain any direct representation of the rich statistical structure of human linguistic experience. Instead, a small number of parameters must be set to determine a particular syntax and phonology, and the lexical entries for words need contain only a determinate pronunciation, a small amount of morphosyntactic information, and a pointer into some symbolically-represented (but non-linguistic) conceptual space. On another view, effective modeling of human language use requires a considerable body of (implicit) knowledge about the relative frequencies of permitted alternatives at all levels of analysis. This second view is once again respectable and even ascendent. It is unlikely that the last word in this discussion will be spoken during our lifetimes, but we can count on seeing a productive and empirically-grounded exploration of the issues during the next decade.

References

1. ACL: 1989, 'ACL Data Collection Initiative Announcement', *The Finite String* **15**.
2. Bahl, L.B., Brown, P.F., de Souza, P.V., and Mercer, R.L.: 1990, 'A Tree-Based Statistical Language Model for Natural Language Speech Recognition'. In Waibel, A., and Lee, K.-F. (eds.), *Readings in Speech Recognition*, San Mateo, CA: Morgan Kaufman.
3. Brill, E., Magerman, D., Marcus, M., and Santorini, B.: 1990, 'Deducing Linguistic Structure from the Statistics of Large Corpora'. In *Proceedings of the DARPA Speech and Natural Language Workshop*, New York: Morgan Kaufman.
4. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Lai, J.C., Mercer, R.L.: 1990, 'An Estimate of an Upper Bound for the Entropy of English'. Ms.
5. Brown, P.F., Cocke J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., and Roosin, P.S.: 1990, 'A Statistical Approach to Machine Translation'. *Computational Linguistics* **16**, 79-85.
6. Chitrao, M., and Grishman, R.: 1990, 'Statistical Parsing of Messages'. In *Proceedings of DARPA Speech and Natural Language Processing Workshop*. New York: Morgan Kaufman.
7. Chomsky, N.: 1957, *Syntactic Structures*. The Hague: Mouton.

8. Choueka, Y.: 1988, 'Looking for Needles in a Haystack: Or, Locating Interesting Collocational Expressions in Large Textual Databases. In *Proceedings of the RIAO88 Conference on User-Oriented Content-Based Text and Image Handling.* Cambridge, MA.
9. Church, K.W.: 1988, 'A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text'. In *Proceedings of the Second ACL Conference on Applied Natural Language Processing.* Austin, Texas.
10. Church, K.W. and Hanks, P.: 1990, 'Word Association Norms, Mutual Information and Lexicography'. *Computational Linguistics* **16**, 22-29.
11. Church, K.W., Hanks, P., and Hindle, D.: forthcoming, 'Using Statistics in Lexical Analysis'. In Zernik, V., ed. *Lexical Acquisition: Using On-line Resources to Build a Lexicon.*
12. Dagan, I., and Itai, A.: 1991 'A Statistical Filter for Resolving Pronoun References'. In *Proceedings of the 29th Meeting of the ACL*, Berkeley.
13. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., and Harshman, R.: 1990, 'Indexing by Latent Semantic Analysis'. *Journal of the American Society for Information Science.*
14. De Marcken, C.G.: 1990, 'Parsing the LOB Corpus'. In *Proceedings of the 28th Annual Meeting of the ACL*, Pittsburgh, PA, 243-251.
15. DeRose, S.J.: 1988, 'Grammatical Category Disambiguation by Statistical Optimization'. *Computational Linguistics* 14, 31-39.
16. Fillmore, C.J., and Atkins, B.T.: forthcoming, 'Toward a Frame-Based Lexicon: the Semantics of RISK and Its Neighbors'. In Lehrer, A., and Kittay, E. (eds.) *Papers in Lexical Semantics.*
17. Gale, W.A. and Church, K.W.: 1990, 'Poor Estimates of Context Are Worse than None'. In *Proceedings of the DARPA Speech and Natural Language Workshop*, June 1990.
18. Hanson, S.J. and Kegl, J.: 1987, 'PARSNIP: A Connectionist Network That Learns Natural Language Grammar from Exposure to Natural Language Sentences'. In *Proceedings of the Cognitive Science Society*, Seattle, WA, 106-119.
19. Hindle, D.: 1990, 'Noun Classification from Predicate-Argument Structures'. In *Proceedings of the 28th Annual Meeting of the ACL*, Pittsburgh, PA, 268-275.
20. Hindle, D. and Rooth., M.: 1990, 'Structural Ambiguity and Lexical Relations'. In *Proceedings of the DARPA Speech and Natural Language Workshop.* June 1990.
21. Jelinek, F.: 1990, 'Self-Organized Language Modeling for Speech Recognition'. In Waibel, A., and Lee, K.-F. (eds.), *Readings in Speech Recognition*, San Mateo, CA: Morgan Kaufman.
22. Jelinek, F., Lafferty, J.D., and Mercer, R.L.: 1990, *Basic Methods of Probabilistic Context Free Grammars.* Yorktown Heights: IBM RC 16374 (#72684).
23. Jelinek, F. and Mercer, R.: 1980, 'Interpolated Estimation of Markov Source Parameters from Sparse Data'. In *Proceedings of the Workshop on Pattern Recognition in Practice.* Amsterdam: North-Holland.
24. Johansson, S., Atwell, E., Garside, R., and Leech, G.: 1986, *The Tagged LOB Corpus: User's Manual.* Bergen: Norwegian Computing Centre for the Humanities.
25. Kernighan, M.D., Church, K.W., and Gale, W.A.: 1990, 'A Spelling Corrector Based on Error Frequencies'. In *Proceedings of the Thirteenth International Conference on Computational Linguistics.*
26. Kroch, A.: 1989 'Function and Grammar in the History of English: Periphrastic *Do*'. In Fasold, R., and Schiffrin, D. (eds.), *Language Change and Variation.* Amsterdam and Philadelphia: John Benjamins.

27. Kucera, H. and Francis, W.N.: 1967, *Computational Analysis of Present-Day American English*. Providence: Brown University Press.
28. Liberman, M.: 1989, 'Text on Tap: the ACL/DCI'. In *Proceedings of the DARPA Speech and Natural Language Workshop*, October 1989. San Mateo, CA.: Morgan Kaufmann.
29. Miller, G.A., and Chomsky, N.: 1963, 'Finitary Models of Language Users'. In Luce, R.D., Bush, R.R., and Galanter, E. (eds.), *Handbook of Mathematical Psychology. Vol. 2*, 419-492. Wiley.
30. Partee, B., Ter Meulen, A., and Wall, W.: 1990, *Mathematical Methods in Linguistics*. Dordrecht: Reidel.
31. Shannon, C.: 1951, 'Prediction and Entropy of Printed English', *Bell Systems Technical Journal* **30**, 50-64.
32. Sinclair, J.M. (ed.): 1987, *Looking Up: An Account of the COBUILD Project in Lexical Computing*. London and Glasgow: Collins.
33. Smadja, F.: 1989, 'Macrocoding the Lexicon with Co-occurrence Knowledge'. In *Proceedings of the First International Lexical Acquisition Workshop*, IJCAI, Detroit, August 1989.
34. Smadja, F. and McKeown, K.: 1990, 'Automatically Extracting and Representing Collocations for Language Generation'. In *Proceedings of the 28th Annual Meeting of the ACL*, Pittsburgh, PA, 252-259.
35. Srihari, S.N.: 1984, *Computer Text Recognition and Error Correction*. IEEE Computer Society Press.
36. Walker, D.: 1989, 'Developing Lexical Resources'. In *Proceedings of the 5th Annual Conference of the UW Centre for the New Oxford English Dictionary*, Waterloo, Ontario.