University of Toronto, Department of Computer Science
CSC 2501F—Computational Linguistics, Fall 2021

Reading assignment 3

Due date: Electronically by 12:10, Friday 15th October 2021.
Late write-ups will not be accepted without documentation of a medical or other emergency.
This assignment is worth 5% of your final grade.

What to read

Fernando Pereira, "Formal grammar and information theory: together again?," *Phil. Trans. R. Soc. Lond.* A, 358, 2000, 1239–1253.

What to write

Write a *brief* summary of the paper's argumentation, with a critical assessment of its merits.

Some points to consider:

- How does our discussion of grammaticality at the beginning of this term bear upon the claims made in this paper?

General requirements: Your write-up should be typed, using 12-point font and 1.5-line spacing; it should fit on one to two sides of a sheet of paper. Submit using the `teach.cs submit` command:

```
$ submit -c csc2501h -a Essay3 essay3.pdf
```

THE ROYAL SOCIETY

Formal grammar and information theory: together again?

By Fernando Pereira

AT&T Laboratories Research, A247, Shannon Laboratory, 180 Park Avenue, Florham Park, NJ 07932-0971, USA (pereira@research.att.com)

In the last 40 years, research on models of spoken and written language has been split between two seemingly irreconcilable traditions: formal linguistics in the Chomsky tradition, and information theory in the Shannon tradition. Zellig Harris had advocated a close alliance between grammatical and information-theoretic principles in the analysis of natural language, and early formal-language theory provided another strong link between information theory and linguistics. Nevertheless, in most research on language and computation, grammatical and information-theoretic approaches had moved far apart.

Today, after many years on the defensive, the information-theoretic approach has gained new strength and achieved practical successes in speech recognition, information retrieval, and, increasingly, in language analysis and machine translation. The exponential increase in the speed and storage capacity of computers is the proximate cause of these engineering successes, allowing the automatic estimation of the parameters of probabilistic models of language by counting occurrences of linguistic events in very large bodies of text and speech. However, I will argue that information-theoretic and computational ideas are also playing an increasing role in the scientific understanding of language, and will help bring together formal-linguistic and information-theoretic perspectives.

Keywords: formal linguistics; information theory; machine learning

1. The great divide

In the last 40 years, research on models of spoken and written language has been split between two seemingly irreconcilable points of view: formal linguistics in the Chomsky tradition, and information theory in the Shannon tradition. The famous quote of Chomsky (1957) signals the beginning of the split.

> (1) Colourless green ideas sleep furiously.
>
> (2) Furiously sleep ideas green colourless.
>
> ...It is fair to assume that neither sentence (1) nor (2) (nor indeed any part of these sentences) has ever occurred in an English discourse. Hence, in any statistical model for grammaticalness, these sentences will be ruled out on identical grounds as equally 'remote' from English. Yet (1), though nonsensical, is grammatical, while (2) is not.

Before and after the split, Zellig Harris had advocated a close alliance between grammatical and information-theoretic principles in the analysis of natural language (Harris 1951, 1991). Early formal-language theory provided another strong link between information theory and linguistics. Nevertheless, in most research on language and computation, those bridges were lost in an urge to take sides that was as much personal and ideological as scientific.

Today, after many years on the defensive, the information-theoretic approach is again thriving and has led to practical successes in speech recognition, information retrieval, and, increasingly, in language analysis and machine translation. The exponential increase in the speed and storage capacity of computers is the proximate cause of these successes, allowing the automatic estimation of the parameters of computational models of language by counting occurrences of linguistic events in very large bodies of text and speech. However, vastly increased computer power would be irrelevant if automatically derived models or linguistic data were not able to *generalize* to unseen data. I will argue below that progress in the design and analysis of such models is not only playing a central role in those practical advances, but also carries the promise of fundamentally deeper understanding of information-theoretic and computational-complexity constraints on language acquisition.

2. Harris's program

The ascent of Chomskian generative linguistics in the early 1960s swept the focus of attention away from distributional views of language, especially those based on the earlier structuralist tradition. In that tradition, Zellig Harris developed what is probably the best-articulated proposal for a marriage of linguistics and information theory. This proposal involves four main so-called *constraints* (Harris 1988) as follows.

Partial order '…for each word…there are zero or more classes of words, called its arguments, such that the given word will not occur in a sentence unless one word…of each of its argument classes is present.'

There's a strong similarity between the argument class information for a word as suggested by Harris and its type in categorial grammar, or subcategorization frames in other linguistic formalisms. However, traditional categorial grammar (Lambek 1958) conflates function–argument relationships and linear order, whereas Harris factors out linear order explicitly. It is only more recently that categorial grammar has acquired the technical means to investigate such factorizations (Morrill 1994; Moortgat 1995). It then becomes clear that Harris's partial order may be formalized as the partial order among set-theoretic function types. However, unlike modern categorial grammar, Harris's partial order constraint specifies only the basic configurations corresponding to elementary clauses, while complex clauses are a result of applying another constraint, reduction, to several elementary clauses.

Likelihood '…each word has a particular and roughly stable likelihood of occurring as argument, or operator, with a given other word, though there are many cases of uncertainty, disagreement among speakers, and change through time.'

Using current terminology, one might interpret the likelihood constraint as a probabilistic version of selectional restrictions. However, Harris makes a sharp distinction between general language, in which likelihoods for fillers of argument positions represent tendencies, and technical sublanguages, in which there are hard constraints on argument fillers, and which, thus, correspond more closely to the usual notion of selectional restriction.

Reduction 'It consists, for each language, of a few specifiable types of reduction... what is reduced is the high-likelihood... material...; an example is zeroing the repeated corresponding words under *and*.'

The reduction constraint tries to account both for morphological processes like contraction, and for processes that combine elementary clauses into complex clauses, such as relativization, subordination and coordination. In each case, Harris claims that high-likelihood material may be elided, although it would seem that additional constraints on reduction may be necessary. Furthermore, connections between reduction-based and transformational analyses (Harris 1965; Chomsky 1965) suggest the possibility of modelling string distributions as the overt projection of a hidden generative process involving operator-argument structures subject to the likelihood constraint and transformations. Recent work linking transformational and categorial approaches to syntax makes this possibility especially intriguing (Stabler 1997; Cornell 1997).

Linearization 'Since the relation that makes sentences out of words is a partial order, while speech is linear, a linear projection is involved from the start.'

Harris's theory left this step rather underspecified. Chomskian transformational grammar can be seen as an effort to fill in the gap with specific mechanisms of sentence generation that could be tested against native speaker grammaticality judgements.

Thus, linguistic events involve the generation of basic configurations—unordered simple clauses—whose structure is determined by the partial order constraint and whose distribution follows the probabilities associated with the likelihood constraint. Those probabilities also govern the application of reduction—compression—to individual configurations or sets of linked configurations. Finally, linearization yields the observable aspects of the event. As I will discuss in § 7, though, the likelihood constraint as stated by Harris, or its current versions, leaves out dependences on the broader discourse context that strongly affect the likelihoods of linguistic events.

For the present discussion, the most important feature of Harris's constraints is how they explicitly link linguistic structure with distributional regularities involving the relative frequencies of different structural configurations. In particular, Harris suggested how the structural and distributional regularities could work together to support language acquisition and use:

> ... when only a small percentage of all possible sound-sequences actually occurs in utterances, one can identify the boundaries of words, and their relative likelihoods, from their sentential environment....

3. Generalization

While Harris discussed the functional role of distributional regularities in language, he proposed no specific mechanisms by which language users could take advantage of those regularities in language acquisition and use. In particular, it is not obvious that language users can acquire stable distributional information, let alone the lexical and grammatical information required by the partial-order, reduction and linearization constraints, from the limited evidence that is available to them from their linguistic environment. This question created a great opening for Chomsky's rationalist critique of empiricist and structuralist linguistics, of which the 'green-ideas' quote above is an early instance.

Chomsky concluded that sentences (1) and (2) are equally unlikely from the observation that neither sentence or 'part' thereof would have occurred previously (Abney 1996). From this observation, he argued that any statistical model based on the frequencies of word sequences would have to assign equal, zero, probabilities to both sentences. But this relies on the unstated assumption that any probabilistic model necessarily assigns zero probability to unseen events. Indeed, this would be the case if the model probability estimates were just the relative frequencies of observed events (the *maximum-likelihood* estimator). But we now understand that this naive method badly *overfits* the training data.

The problem of overfitting is tightly connected with the question of how a learner can generalize from a finite training sample. The canonical example is that of fitting a polynomial to observations. Given a finite set of observed values of a dependent random variable Y for distinct values of the independent variable X, we seek a hypothesis for the functional dependency of Y on X. Now, any such set of observations can be fitted exactly by a polynomial of high-enough degree. But that curve will typically be a poor predictor of a new observation because it exactly matches the peculiarities of the training sample. To avoid this, one usually *smooths* the data, using a lower-degree polynomial that may not fit the training data exactly but that will be less dependent on the vagaries of the sample. Similarly, smoothing methods can be used in probability models to assign some probability mass to unseen events (Jelinek & Mercer 1980). In fact, one of the earliest such methods, due to Turing and Good (1953), had been published before Chomsky's attack on empiricism, and has since been used to good effect in statistical models of language (Katz 1987).

The use of smoothing and other forms of *regularization* to constrain the form of statistical models and ensure better generalization to unseen data is an instance of a central theme in statistical learning theory, that of the *sample complexity* relationship between training sample size, model complexity and generalization ability of the model. Typical theoretical results in this area give probabilistic bounds on the generalization error of a model as a function of model error on training data, sample size, model complexity, and margin of error (Vapnik 1995). In qualitative terms, the gap between test and training error—a measure of overfitting—grows with model complexity for a fixed training sample size, and decreases with sample size for a fixed model complexity.

To quantify the trade-off between training-set accuracy, generalization to new data and constraints on the model, we need a rigorous measure of model complexity. In the polynomial example, the usual intuition is that complexity is measured by the degree of the polynomial (the number of tunable coefficients in the model), but intuitions are harder to come by for model classes without a simple parametric form. Furthermore, even in the polynomial case, the common-sense complexity measure can be misleading, because certain approaches to polynomial fitting yield much smaller model complexity and, thus, better generalization ability (Vapnik 1995). The definition of model complexity is also intimately tied to the learning setting; for instance, whether one assumes that the data have a distribution of known form but unknown parameters (as is usually done in statistics), or one takes a *distribution-free* view, in which the data are distributed according to an unknown (but fixed between training and test) distribution (Valiant 1984), or even one assumes an *on-line* setting, in which the goal is to do the best possible prediction on a fixed sequence incrementally generated by the environment (Littlestone & Warmuth 1994; Freund & Schapire 1997). A crucial idea from the distribution-free setting is that model complexity can be measured, even for an infinite model class, by combinatorial quantities such as the *Vapnik–Chervonenkis (VC) dimension* (Vapnik & Chervonenkis 1971), which, roughly speaking, gives the order of a polynomial upper bound on the number of distinctions that can be made between samples by models in the class, as a function of sample size.

Returning to the debate between empiricism and rationalism, the relationships between model complexity, sample size and overfitting developed in learning theory may help clarify the famous argument from *poverty of the stimulus* (APS). Reacting to empiricist and especially behaviourist theories, Chomsky and others have argued that general-purpose learning abilities are not sufficient to explain children's acquisition of their native language from the (according to them) very limited linguistic experience that is available to the learner. In particular, they claimed that linguistic experience does not provide negative examples of grammaticality, making the learner's task that much harder. Therefore, they conclude, a specialized innate language faculty must be involved. The 'green-ideas' example is an early instance of the same argument, asserting that statistical procedures alone cannot acquire a model of grammaticality from the data available to the learner.

The APS does not just require restrictions on model classes to ensure effective generalization from finite data, which would be unobjectionable from a learning-theoretic viewpoint. In its usual form, the APS also claims that only a learning mechanism developed specifically for language could generalize well from limited linguistic experience. The flaw in this argument is that it implicitly assumes that the only constraints on a learner are those arising from particular *representations* of the learner's knowledge, whereas we now know that the informational difficulty of learning problems can be characterized by purely combinatorial, representation-independent, means. Statistical learning theory gives us the tools to compute empirically testable lower bounds on sample sizes that would guarantee learnability for given model classes, although such bounds can be very pessimistic unless they take into account constraints on the model search procedure as well. Nevertheless, it is unlikely that the debate over the APS can become empirically grounded without taking into account such calculations, since the stimuli that APS supporters claimed to be missing are actually present with significant frequency (Pullum 1996).

The APS reached an extreme form with Chomsky's principles-and-parameters theory, according to which learnability requires that the set of possible natural languages be generated by the settings of a finite set of finitely valued parameters (Chomsky 1986, p. 149). But this extreme constraint is neither necessary, since infinite model classes of finite VC dimension are learnable from an information-theoretic point of view, nor sufficient, because even finite classes may not be *efficiently* learnable, that is, the search for a model with good generalization may be computationally intractable† even though the information is, in principle, available (Kearns & Valiant 1994).

4. Hidden variables

Early empiricist theories of linguistic behaviour made themselves easy targets of critiques like that of Chomsky (1959) by denying a significant role for the internal, unobservable, state of the language user. Thus, in a Markov model of language, all the state information would be represented by the externally observable sequence of past linguistic behaviour. However, even in this case, the empiricist position had been somewhat oversimplified by its critics. If we consider a language user that updates its expectations and probable responses according to statistics collected from its past experience, those expectations and response propensities, however represented, are a part of the user state that is not directly available for observation. Furthermore, the behaviour of language users may give valuable information about the power of their experience-encoding mechanisms. For instance, a language user that maintains statistics over pairs of consecutive words only (bigram statistics) might be less effective in anticipating and reacting appropriately to the next word than a user that keeps statistics over longer word sequences; in other words, the bigram model may have higher entropy. This example, related to finite-state text compression, may seem simplistic from the point of view of linguistics, but it is a convenient test-bed for ideas in statistical modelling and constraints on model structure, and introduces the idea of a hidden modelling state in a very simple form.

Hidden *random* variables in a language user's state—or, rather, statistics involving their joint values—represent the user's uncertainty about the interpretation of, and best response to, events observed so far. Such uncertainty may not just be over the interpretation of a particular course of events, but also over which particular model in a class of models is a best compromise between fitting the experience so far and generalizing to new experience. When the best choice of model is uncertain, Bayesian *model averaging* (Willems *et al.* 1995) can be used to combine the predictions of different candidate models according to the language user's degree of belief in them, as measured by their past success. Model averaging is, thus, a way for learners to hedge their bets on particular grammars, in which the initial bets represent a *prior* belief on particular grammars and are updated according to a regularizing procedure that balances fit to past experience with predictive power over new experience. The prior distribution on grammars can be seen as a form of innate knowledge that implicitly biases the learner towards 'better'—for instance, less complex—grammars. In particular, any infinite class of grammars can be given a *universal* prior based on the number of bits needed to encode members of the class, which favours the least complex grammars compatible with the data (Solomonoff 1964; Horning 1969). However, those results did not provide a way of quantifying the relationship between a prior over grammars, training sample size and generalization power, and in any case seems to have been ignored by those interested in language acquisition and the APS. Recent advances in statistical learning theory (McAllester 1999) may provide new theoretical impetus to that research direction, since they show that a prior

† I use 'intractable' in this paper in the usual sense from theory of computation of a problem that has been proven to belong to one of the standard classes believed to require more than polynomial time on a deterministic sequential computer, for instance the NP-hard problems.

over models can play an analogous regularizing role to a combinatorial complexity measure.

The other role for hidden variables, capturing uncertainty in the interpretation of particular experience, becomes especially interesting in modelling ambiguity. For example, going back to Harris's theory, each of the constraints involves covert choices by the language user: assignment of types—positions in the partial order—to lexical items; lexical choice according to selection probabilities; reduction choices according to the distributional statistics of predictability; and linearization choices. More generally, any model of language that appeals to non-observables, for instance any model that assigns syntactic analyses, requires hidden variables.

Hidden variables representing uncertainty of interpretation can also be used to create *factored* models of joint distributions that have far fewer parameters to estimate, and are, thus, easier to learn, than models of the full joint distribution. As a very simple but useful example, we may approximate the conditional probability $p(x, y)$ of occurrence of two words x and y in a given configuration as

$$p(x, y) = p(x) \sum_c p(y \mid c) p(c \mid x),$$

where c is a hidden 'class' variable for the associations between x and y in the configuration under study. For a vocabulary of size V and C classes, this model uses $O(CV)$ parameters rather than the $O(N^2)$ parameters of the direct model for the joint distribution, and is thus less prone to overfitting if $C \ll V$. In particular, when $(x, y) = (v_i, v_{i+1})$, we have an *aggregate* bigram model (Saul & Pereira 1997), which is useful for modelling word sequences that include unseen bigrams. With such a model, we can approximate the probability of a string $p(w_1 \cdots w_n)$ by

$$p(w_1 \cdots w_n) = p(w_1) \prod_{i=2}^{n} p(w_i | w_{i-1}).$$

By using this estimate for the probability of a string and an aggregate model with $C = 16$ trained on newspaper text, and by using the expectation–maximization (EM) method (Dempster *et al.* 1977), we find that

$$\frac{p(\text{Colourless green ideas sleep furiously})}{p(\text{Furiously sleep ideas green colourless})} \approx 2 \times 10^5.$$

Thus, a suitably constrained statistical model, even a very simple one, can meet Chomsky's particular challenge.

A plausible and well-defined model of the statistical dependences among the hidden variables is, however, not in general sufficient, since the problem of setting the corresponding conditional probabilities from observable linguistic material is in most cases computationally intractable (Abe & Warmuth 1992). Nevertheless, those intractability results have not precluded significant algorithmic and experimental progress with carefully designed model classes and learning methods, such as EM and variants, especially in speech processing (Baum & Petrie 1966; Baker 1979). In particular, the learning problem is easier in practice if interactions between hidden variables tend to factor via the observed variables.

5. Lexicalized models

Harris's model of dependency and selection is *lexicalized*, in the sense that all postulated relationships are between the words in (precursors for) sentences, rather than the relationships between structures in generative grammar. From the points of view of distributional modelling and machine learning, an important property of lexicalized models is that they anchor analyses in observable co-occurrences between words, rather than in unobservable relationships among hypothetical grammatical structures.† In a probabilistic setting, a way to state this more precisely is that lexicalization makes it easier to factor the interactions between the hidden variables by conditioning on the observed sentence.

Even lexicalized models will involve hidden decisions if they allow ambiguous interpretations. As noted in the previous section, hidden-variable models are computationally difficult to learn from the observable variables alone. An alternative strategy is to constrain the hidden variables by associating sentences with disambiguating information. At one extreme, that information might be a full analysis. In this case, which is very interesting from computational and applications perspectives, recent work has shown that lexicalized probabilistic context-free grammars can be automatically learned that perform, with remarkable accuracy, on novel material (Charniak 1997; Collins 1998). Besides lexicalization, these models factor the sentence-generation process into a sequence of conditionally independent events that reflect such linguistic distinctions as those of head and dependent and of argument and adjunct. That is, the models are in effect lexically based *stochastic generative grammars*, and the conditional independence assumptions on the generation process are a particular kind of Markovian assumption. Crucially, these assumptions apply to the hidden generative decisions, not to the observable utterance, and, thus, allow for analysis ambiguity.

The learning algorithms just discussed need to be given the full correct syntactic analysis of each training example, and are, thus, not realistic models of human language acquisition. One possible direction for reducing the unrealistic amount of supervision required would be to use additional observables correlated with the hidden variables instead, such as prosodic information or perceptual input associated with the content of the linguistic input (Siskind 1996; Roy & Pentland 1999). More generally, we may be able to replace direct supervision with indirect correlations, as I now discuss.

6. The power of correlations

How poor is the stimulus that the language learner exploits to acquire its native language? As I observed above, linguistic experience is not just a string of words, but it is *grounded* in a rich perceptual and motor environment that is likely to provide crucial clues to the acquisition, interpretation and production processes, if for no other reason than for the functional one that much of the linguistic experience

† This property could well make lexicalized models less rather than more palatable to Chomskian linguists, for whom structural relationships are the prime subject of theory. But notice that Chomsky's more recent 'minimalist program' (Chomsky 1995) is much more lexically based than any of his theories since 'Aspects' (Chomsky 1965), in ways that are reminiscent of other lexicalized multistratal theories, in particular lexical-functional grammar (Bresnan 1982), HPSG (Pollard & Sag 1994), and certain varieties of categorial grammar (Morrill 1994; Moortgat 1995; Cornell 1997).

is *about* that non-linguistic environment. However, this points to a fundamental weakness in much of the work discussed so far: both in formal grammar and in most computational models of language, language is taken as a completely autonomous process that can be independently analysed.† Indeed, a simplistic use of information theory suffers from the same problem, in that the basic measures of information content of a signal are intrinsic, rather than relative to the correlations between a signal and events of interest (the meaning(s) of the signal). In particular, Harris's likelihood and reduction constraints appear to ignore the content-carrying function of utterances. Fortunately, information theory provides a ready tool for quantifying *information about* with the notion of *mutual information* (Cover & Thomas 1991), from which a suitable notion of compression relative to side variables of interest can be defined (Tishby *et al.* 1999).

Given the enormous conceptual and technical difficulties of building a comprehensive theory of grounded language processing, treating language as an autonomous system is very tempting. However, there is a weaker form of grounding that can be exploited more readily than physical grounding, namely grounding in a *linguistic* context. Following this path, sentences can be viewed as evidence for other sentences through inference, and the effectiveness of a language processor may be measured by its accuracy in deciding whether a sentence entails another, or whether an answer is appropriate for a question.

Furthermore, there is much empirical evidence that linguistic grounding carries more information than it might seem to at first sight. For instance, all of the most successful information-retrieval systems ignore the order of words and just use the frequencies of words in documents (Salton 1989) in the so-called *bag-of-words* approach. Since similar situations are described in similar ways, simple statistical similarity measures between the word distributions in documents and queries are effective in retrieving documents relevant to a given query. In the same way, word senses can be automatically disambiguated by measuring the statistical similarity between the bag of words surrounding an occurrence of the ambiguous word and the bags of words associated with definitions or examples of the different senses of the word (Schütze 1997).

In both information retrieval and sense disambiguation, bag-of-words techniques are successful because of the underlying coherence of purposeful language, at syntactic, semantic, and discourse levels. The *one-sense-per-discourse* principle (Gale *et al.* 1992) captures a particular form of this coherence. For example, the co-occurrence of the words 'stocks', 'bonds' and 'bank' in the same passage is potentially indicative of a financial subject matter, and thus tends to disambiguate those word occurrences, reducing the likelihood that the 'bank' is a river bank, that the 'bonds' are chemical bonds, or that the 'stocks' are an ancient punishment device. These correlations, like the correlations between utterances and their physical context, allow a language processor to learn from its linguistic environment with very little or no supervision (Yarowsky 1995), and have suggested new machine-learning settings, such as *co-training* (Blum & Mitchell 1998).

Both lexicalized grammars and bag-of-words models represent statistical associations between words in certain configurations. However, the kinds of associations

† I include under this description all the work on formal semantics of natural language, since logical representations of the meanings of sentences are as unobservable as syntactic analyses, and are thus equally artificial as inputs to a language acquisition process.

represented are rather different. The associations in lexicalized grammars are mediated by a hidden assignment of dependency relationships to pairs of word occurrences in an utterance. Many such assignments are potentially available, leading to great structural ambiguity, as discussed in § 5. In contrast, the associations in bag-of-words models and many other statistical models (for instance, Markov models) are defined over very impoverished but unambiguous overt structures. Furthermore, effective lexicalized models must make very strong statistical-independence assumptions between parts of the underlying structure, thus missing the global coherence correlations that bag-of-words models capture.

7. Local structure and global distribution

Current stochastic lexicalized models, with their lexically determined local correlations, capture much of the information relevant to Harris's partial-order and likelihood constraints. However, unlike Harris, but like dependency grammar and other monostratal grammatical formalisms, they conflate linearization with the argument structure given by the partial-order constraint.

In asserting the 'rough stability' of the likelihood of a given argument of a given operator, Harris implicitly assumed a generative model in which dependents are conditionally independent of the rest of an analysis given the head they depend on. Existing lexicalized models use similar Markovian assumptions, although they typically extend lexical items with additional features, for instance syntactic category (Charniak 1997; Collins 1998). However, Harris's information-theoretic arguments, especially those on reduction, refer to the overall likelihood of a string, which involves the global correlations discussed in the last section. But such global correlations are precisely what the Markovian assumptions in generative models leave out.

Thus, Markovian generative models are not able to model the potential correlations between the senses assigned to occurrences of 'stocks' and 'bonds' in different parts of a paragraph, for example. This problem may be addressed in two main ways. The first is to preserve Markovian assumptions, but to enrich lexical items with features representing alternative global coherence states. For instance, lexical items might be decorated with sense features, and local correlations between those would be used to enforce global coherence. Those features might even be other lexical items, whose co-occurrence with the given items as operators or arguments may disambiguate them. The difficulty with this approach is that it introduces a plethora of hidden variables, leading to a correspondingly harder learning problem. Furthermore, it relies on careful crafting of the hidden variables, for instance in choosing informative sense distinctions. The second approach is to adopt ideas from random fields and factor probabilities instead as products of exponentials of indicator functions for significant local or global *features* (events) (Della Pietra *et al.* 1997; Ratnaparkhi 1997; Abney 1997), which can be built incrementally with 'greedy' algorithms that select the most informative feature at each step.

8. From deciding to understanding

Models based on information-theoretic and machine-learning ideas have been successful in a variety of language-processing tasks, such as speech recognition and information retrieval. A common characteristic of most of those tasks is that what is sought is a decision among a finite set of alternatives, or a ranking of alternatives. For example:

(i) a newswire filter classifies news stories into topics specified by training examples;

(ii) a part-of-speech tagger assigns the most likely tags to the words in a document;

(iii) a Web search engine ranks a set of Web pages according to their relevance to a natural language query;

(iv) a speech recognizer decides among the possible transcriptions of a spoken utterance.

In each case, the task can be formalized as that of learning a mapping from spoken or written material to a choice or ranking among alternatives. As we know from the earlier discussion of generalization, we need to restrict our attention to a class of mappings that can actually be learned from the available data. Computational considerations and experimental evaluation will further narrow the mapping classes under consideration. Finally, a suitable optimization procedure is employed to select a mapping from the class that minimizes some measure of the error on the training set.

A potential weakness of such task-directed learning procedures is that they ignore regularities that are not relevant to the task. Yet, those regularities may be highly informative about other questions. While language may be redundant with respect to any particular question, and a task-oriented learner may benefit greatly from that redundancy, as discussed earlier, it does not follow that language is redundant with respect to the set of all questions that a language user may need to decide. Furthermore, one may reasonably argue that a task-oriented learner does not really 'understand' language, since it can accurately answer only one question, while our intuitions about understanding suggest that a competent language user can accurately answer many questions pertaining to any discourse it processes. For instance, a competent language user should be able to reliably answer 'who did what to whom' questions pertaining to each clause in the discourse.

We are thus drawn to the question of what kinds of learning tasks may involve 'understanding' but do not force us to attack frontally the immense challenges of grounded language processing. Automatically trained machine translation (Brown *et al.* 1990; Alshawi & Douglas, this issue) may be such a task, since translation requires that many questions about a text be accurately answered to produce a correct output. Nevertheless, it is easy to find many other reasonable questions that can be left unanswered while still performing creditably on the task. Indeed, there is no single 'understanding' task, but, rather, a range of tasks whose difficulty can be measured by the uncertainty—information-theoretically, the entropy—of the output in the absence of any information about the input. The objective of a learner is then to acquire a function that can reduce that uncertainty by exploiting the mutual information between inputs and outputs (Tishby & Gorin 1994). Tasks (i)–(iv) above are listed roughly in order of increasing output entropy, with machine translation possibly being even more difficult.

The theoretical representations postulated by formal linguistics (constituent structure, functional and dependency structures, logical form) can also be understood as codified answers to particular kinds of questions pertaining to the text, with their own degrees of information-theoretic difficulty. For instance, different assignments of arguments to thematic roles lead to different correct answers to 'who did what to whom' questions. From this point of view, the task of the learner is to acquire an accurate procedure for deciding whether a simple sentence follows from a discourse, rather than the more traditional tasks of deciding grammaticality or assigning structural descriptions. Structural descriptions will still play an important role in such a theory, but now as proxies for informational relationships between external linguistic events instead of end-products of the theory.

9. Summary

While researchers in information retrieval, statistical pattern recognition, and neural networks kept developing theoretical and experimental approaches to the problem of generalization, that work was ignored by formal linguistics for both cultural and substantive reasons. Among the substantive reasons, possibly the most important was that the models proposed, even if successful in practice, failed to capture the productive, recursive nature of linguistic events.

Recent advances in machine learning and statistical models are starting to supply the missing ingredients. Lexicalized statistical models informed by linguistic notions such as phrase head, argument and adjunct specify how complex linguistic events can be generated and analysed as sequences of elementary decisions. Machine learning suggests how rules for the elementary decisions can be learned from examples of behaviour, and how the learned decision rules generalize to novel linguistic situations. Probabilities can be assigned to complex linguistic events, even novel ones, by using the causal structure of the underlying models to propagate the uncertainty in the elementary decisions.

Such statistical models of local structure are complemented by the models of larger-scale correlations that have been developed in information retrieval and speech recognition. These models have proved to be quite successful in automatically learning how to rank possible answers to a given question, but it is still unclear how they may combine with lexical models in a unified account of the relationship between linguistic structure and statistical distribution.

Furthermore, we have barely touched on the question of what such models may say about human language acquisition. Although statistical learning theory and its computational extensions can help us ask better questions and rule out seductive *non sequiturs*, their quantitative results are still too coarse to significantly narrow the field of possible acquisition mechanisms. However, some of the most successful recent advances in machine learning arose from theoretical analysis (Cortes & Vapnik 1995; Freund & Schapire 1997), and theory is also helping to sharpen our understanding of the power and limitations of informally designed learning algorithms.

All in all, while much remains to be done, we may well be seeing the beginning of a new version of the Harris program, in which computational models constrained by grammatical considerations define broad classes of possible grammars, and information-theoretic principles specify how those models are fitted to actual linguistic data.

I thank Gerald Gazdar and Karen Spärck Jones for their careful reading of this paper and illuminating comments; Ido Dagan, Lillian Lee, Larry Saul, Yves Schabes, Yoram Singer, Amit Singhal and Tali Tishby for the joint research that helped shape these ideas; Yoav Freund, Michael Kearns and Rob Schapire for guidance on learning theory; and Steve Abney, Hiyan Alshawi, Michael Collins, Don Hindle, Mark Johnson, Aravind Joshi, John Lafferty, David McAllester, Glyn Morrill, Michael Moortgat, Hinrich Schütze, Stuart Shieber and Ed Stabler for many conversations on these topics over the years. I am sure that each of them will have good reasons to disagree with some of my arguments and interpretations, but nevertheless their help was invaluable in this effort to reconcile the two rich traditions in the study of language that most of my work derives from.

References

Abe, N. & Warmuth, M. 1992 On the computational complexity of approximating distributions by probabilistic automata. *Machine Learning* **9**, 205–260.

Abney, S. 1996 Statistical methods and linguistics. In *The balancing act* (ed. J. L. Klavans & P. Resnik). Cambridge, MA: MIT Press.

Abney, S. 1997 Stochastic attribute-value grammars. *Comp. Ling.* **23**, 597–618.

Baker, J. K. 1979 Trainable grammars for speech recognition. In *97th Mtg of the Acoustical Society of America* (ed. J. J. Wolf & D. H. Klatt). Cambridge, MA. Acoustical Society of America.

Baum, L. E. & Petrie, T. 1966 Statistical inference for probablistic functions of finite state Markov chains. *Ann. Math. Stat.* **37**, 1554–1563.

Blum, A. & Mitchell, T. 1998 Combining labeled and unlabeled data with co-training. In *Proc. 11th Ann. Conf. on Computational Learning Theory*. New York: ACM.

Bresnan, J. (ed.) 1982 *The mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Brown, P., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L. & Roossin, P. S. 1990 A statistical approach to machine translation. *Comp. Ling.* **16**, 79–85.

Charniak, E. 1997 Statistical parsing with a context-free grammar and word statistics. In *14th Natl Conf. on Artificial Intelligence*, pp. 598–603. Cambridge, MA: AAAI Press/MIT Press.

Chomsky, N. 1957 *Syntactic structures*. The Hague: Mouton.

Chomsky, N. 1959 Review of Skinner's 'verbal behavior'. *Language* **35**, 26–58.

Chomsky, N. 1965 *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1986 *Knowledge of language: its nature, origin, and use*. New York: Praeger.

Chomsky, N. 1995 *The minimalist program*. Cambridge, MA: MIT Press.

Collins, M. 1998 Head-driven statistical models for natural language parsing. PhD thesis, University of Pennsylvania, USA.

Cornell, T. 1997 A type-logical perspective on minimalist derivations. In *Formal Grammar '97* (ed. G.-J. van Kruijff & R. Oehrle). Aix-en-Provence.

Cortes, C. & Vapnik, V. N. 1995 Support vector networks. *Machine Learning* **20**, 273–297.

Cover, T. M. & Thomas, J. A. 1991 *Elements of information theory*. Wiley.

Della Pietra, S. A., Della Pietra, V. J. & Lafferty, J. D. 1997 Inducing features of random fields. *IEEE Trans. Pattern Analysis Machine Intel.* **19**, 380–393.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the *EM* algorithm. *J. R. Stat. Soc.* **39**, 1–38.

Freund, Y. & Schapire, R. 1997 A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comp. Syst. Sci.* **55**, 119–139.

Gale, W. A., Church, K. W. & Yarowsky, D. 1992 One sense per discourse. In *Proc. 4th DARPA Speech and Natural Language Workshop*, pp. 233–237. San Francisco, CA: Morgan Kaufmann.

Good, I. J. 1953 The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–264.

Harris, Z. S. 1951 *Structural linguistics.* University of Chicago Press.

Harris, Z. S. 1965 *String analysis of sentence structure.* The Hague: Mouton.

Harris, Z. S. 1988 *Language and information.* New York: Columbia University Press.

Harris, Z. S. 1991 *A theory of language and information: a mathematical approach.* Oxford: Clarendon.

Horning, J. J. 1969 A study of grammatical inference. PhD thesis, Stanford University, USA.

Jelinek, F. & Mercer, R. L. 1980 Interpolated estimation of Markov source parameters from sparse data. In *Proc. Workshop on Pattern Recognition in Practice.* Amsterdam: North Holland.

Katz, S. M. 1987 Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Trans. Acoust. Speech Sig. Proc.* **35**, 400–401.

Kearns, M. J. & Valiant, L. G. 1994 Cryptographic limitations on learning Boolean formulae and finite automata. *J. ACM* **41**, 67–95.

Lambek, J. 1958 The mathematics of sentence structure. *Am. Math. Mon.* **65**, 154–170.

Littlestone, N. & Warmuth, M. 1994 The weighted majority algorithm. *Informat. Comput.* **108**, 212–261.

McAllester, D. A. 1999 PAC-Bayesian model averaging. In *Proc. 12th Ann. Conf. on Computational Learning Theory*, pp. 164–170. New York: ACM.

Moortgat, M. 1995 Multimodal linguistic inference. *Bull. Interest Group Pure Appl. Logics* **3**, 371–401.

Morrill, G. V. 1994 *Type logical grammar: categorial logic of signs.* Dordrecht: Kluwer.

Pollard, C. & Sag, I. A. 1994 *Head-driven phrase structure grammar.* University of Chicago Press.

Pullum, G. K. 1996 Learnability, hyperlearning, and the poverty of the stimulus. In *Proc. 22nd Ann. Mtg General Session and Parasession on the Role of Learnability in Grammatical Theory* (ed. J. Johnson, M. L. Juge & J. L. Moxley), pp. 498–513. Berkeley, CA: Berkeley Linguistics Society.

Ratnaparkhi, A. 1997 A linear observed time statistical parser based on maximum entropy models. In *2nd Conf. Empirical Methods in Natural Language Processing (EMNLP-2)* (ed. C. Cardie & R. Weischedel). Somerset, NJ: Association for Computational Linguistics.

Roy, D. & Pentland, A. 1999 Learning words from natural audio-visual input. In *Int. Conf. on Spoken Language Processing*, vol. 4, pp. 1279–1283. Sydney, Australia: Australian Speech Science and Technology Association.

Salton, G. 1989 *Automatic text processing—the transformation, analysis and retrieval of information by computer.* Reading, MA: Addison-Wesley.

Saul, L. & Pereira, F. 1997 Aggregate and mixed-order Markov models for statistical language processing. In *Proc. 2nd Conf. on Empirical Methods in Natural Language Processing* (ed. C. Cardie & R. Weischedel), pp. 81–89. Somerset, NJ: Association for Computational Linguistics. (Distributed by Morgan Kaufmann, San Francisco, CA.)

Schütze, H. 1997 *Ambiguity resolution in language learning: computational and cognitive models.* Stanford, CA: CSLI.

Siskind, J. M. 1996 A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* **61**, 39–91.

Solomonoff, R. J. 1964 A formal theory of inductive inference. *Informat. Control* **7**, 1–22, 224–254.

Stabler, E. 1997 Derivational minimalism. In *Logical aspects of computational linguistics* (ed. C. Retoré), pp. 68–95. Springer.

Tishby, N. & Gorin, A. 1994 Algebraic learning of statistical associations. *Comp. Speech Lang.* **8**, 51–78.

Tishby, N., Pereira, F. & Bialek, W. 1999 Extracting relevant bits: the information bottleneck method. In *Proc. 37th Allerton Conf. on Communication, Control and Computing* (ed. B. Hajek & R. S. Sreenivas). Urbana, IL: University of Illinois.

Valiant, L. G. 1984 A theory of the learnable. *Commun. ACM* **27**, 1134–1142.

Vapnik, V. N. 1995 *The nature of statistical learning theory.* Springer.

Vapnik, V. N. & Chervonenkis, A. Y. 1971 On the uniform convergence of relative frequencies of events to their probabilities. *Theory Prob. Appl.* **16**, 264–280.

Willems, F., Shtarkov, Y. & Tjalkens, T. 1995 The context tree weighting method: basic properties. *IEEE Trans. Informat. Theory* **41**, 653–664.

Yarowsky, D. 1995 Unsupervised word sense disambiguation rivaling supervised methods. In *33rd Ann. Mtg Association for Computational Linguistics*, pp. 189–196. Somerset, NJ: Association for Computational Linguistics.

Discussion

J. Coleman (*University of Oxford, UK*). What is the status of statistical information in natural language processing systems—does it compensate for gaps in our theoretical knowledge or model humans' behaviour in the face of uncertainty?

F. Pereira. The uncertainty may be about how much linguistic knowledge is shared between interlocutors. Such variation within populations is often modelled as a statistical distribution. Furthermore, this variation is what drives language change in speech communities. See the work by Anthony Kroch and colleagues at the University of Pennsylvania. But there is also increasing evidence from psycholinguists (see work of Michael Spivey and colleagues at Cornell University) that uncertainty is explicitly represented (e.g. by ranging activation levels in human language processing).

P. A. Taylor (*University of Edinburgh, UK*). Have you tested nonsensical and ungrammatical sentences on a smoothed language model to confirm that they were assigned appropriately different probabilities?

F. Pereira. No, but others have confirmed that such sentences are indeed assigned positive and differing probabilities. *Note added in proof.* The published version of the paper now discusses a simple experiment answering this question.

K. Ahmad (*University of Surrey, UK*). Did Zellig Harris not apply his model primarily to specialized languages?

F. Pereira. Harris made too strong a distinction between specialized, scientific language and more general language, arguing, for instance, that selection restrictions are inviolable in the former but not in the latter, but this is probably a difference in the degree of likelihood of such restrictions holding.