[6] S. Finch and N. Chater, "Bootstrapping syntactic categories using statistical methods," in *Background and Experiments in Machine Learning of Natural Language* (D. Daelemans, W. & Powers, ed.), (Tilburg, NL.), pp. 229–236, ITK, 1992.

[7] D. Caplan, *Neurolinguistics and Linguistic Aphasiology*. Cambridge: Cambridge University Press, 1987.

[8] N. Geschwind, "The paradoxical position of kurt goldstein in the history of aphasia," *Cortex*, vol. 1, pp. 214–224, 1964.

[9] M. F. Garrett, "Syntactic processes in sentence production," in *New Approaches to Language Mechanisms* (R. Wales and E. Walker, eds.), Amsterdam: North-Holland, 1976.

[10] M. Garrett, "The organization of processing structure for language production," in *Biological Perspectives on Language* (D. Caplan, A. Lecours, and A. Smith, eds.), Cambridge, Mass.: MIT Press, 1984.

[11] S. Abney, "Functional elements and licensing." presented to GLOW, Gerona, Spain, April 1986.

[12] T. C. Smith and I. H. Witten, "Language inference from function words," Working Paper Series 1170-487X-1993/3, Department of Computer Science, University of Waikato, Hamilton, New Zealand, August 1993.

[13] M. Cleveland, "Language inference from a closed class," honours thesis, University of Waikato, Hamilton, New Zealand, 1994.

[14] T. C. Smith, I. H. Witten, J. Cleary, and S. Legg, "Objective evaluation of inferred context-free grammars," in *Proceedings of the Australasian Computer Conference-94*, (Brisbane, Australia), November 1994.

category	elements		
cw_{41}	pulled	sent	drew
	wrong	formed	asked
	visible	returned	short
	used	closed	
cw_{44}	certainly	merely	entirely
	already	apparently	sometimes
	really	nearly	hardly
cw_{57}	doing	beginning	able
	coming	next	began
	feeling	looking	having
	going		
cw_{58}	miles	circumstances	pounds
	clothes	hours	arms
	feet	neighbours	thoughts
	horses	trees	features
	lips	days	others
	sort	hands	minutes
	things	times	people
	sheep	women	years
	words	men	

Table 6: Some open-class lexical categories for *Far From the Madding Crowd*

categories we obtained with those inferred from other statistical techniques, particularly ones that produce nonstandard categories, such as Finch and Chater's. However, we can argue that the algorithms involved are at least tractable for the complete vocabularies of even very large corpora, and that they have been shown to be effective when they are incorporated into other language learning systems.

Acknowledgements

References

[1] E. Charniak, *Statistical language learning*. Massachusetts: MIT Press, 1993.

[2] J. A. Feldman, "Some decidability results on grammatical inference and complexity," AI Memo 93.1, Computer Science Dept., Stanford University, Stanford, California, 1970.

[3] E. Brill, "A simple rule-based part of speech tagger," in *Proceedings of the Third Conference on Applied Computational Linguistics*, (Trento, Italy), 1992.

[4] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun, "A practical part-of-speech tagger," in *Proceedings of the 3rd Conference on Applied Natural Language (ACL)*, (Trento, Italy), 1993.

[5] E. Brill and M. Marcus, "Automatically acquiring phrase structure using distributional analysis," in *Darpa Workshop on Speech and Natural Language*, (Harriman, N. Y.), 1992.

$$cw(fw_0, fw_7, 1, 3) = \{\text{tiny}\}$$
$$cw(fw_0, fw_7, 2, 3) = \{\text{bird}\}$$
$$cw(fw_0, fw_7, 3, 3) = \{\text{sat}\}$$
$$cw(fw_0, fw_\phi, 1, 1) = \{\text{tree}\}.$$

For example, *bird* is assigned to the set of words appearing in second position of a phrase of length 3 headed by a word from the fw_0 category and followed by a phrase headed by a word from fw_7. Similarly, *tree* is assigned to an open-class category for words appearing in the first position of a phrase of length 1 headed by fw_0 and followed by the end of a sentence (i.e. the empty phrase). As each sentence is processed, previously unseen content words are added to existing sets, or new categories are created for them. A word can be assigned to several categories, though duplicates are removed within each category.

Open-class category generalization

When applied to *Far From The Madding Crowd*, this procedure creates 15,534 initial categories. Each is subsequently compared against all others in the same manner as the first-order successors for closed-class words were compared. That is, the strength of the association between two categories is determined by the probability that the sets have an intersection of the size exhibited. The larger the intersection, the more likely it is that the categories share the same lexical function. Probabilities are calculated for all pairs before any are combined, and amalgamation is performed in a single pass. Once again, no provision is made to prevent a word from occupying several categories.

Table 6 shows some of the 61 final open-class categories derived using this technique. Category cw_{44} exemplifies a fairly sound collection of adverbs, and cw_{41} and cw_{57} are reasonably consistent sets of past tense and present participle verbs respectively. Category cw_{58} includes many of the plural nouns from *Far From The Madding Crowd*. These groupings represent some of the more coherent open-class categories; however, they do not demonstrate complete collections of the classic grammatical forms they exemplify. For example, most of the present participle verbs used in Hardy's novel are found in groupings not listed here, often mixed in with words from a variety of standard syntactic categories. Of the 61 categories, 58 contain fewer than 170 words, each of which tends toward a particular grammatical class. Unfortunately the three largest sets contain over 3000 words and do not submit to characterization under traditional syntactic forms. In general, the larger the group the more difficult it is to interpret using standard grammatical terminology. A similar effect occurred in Finch and Chater's system where 60% of the vocabulary ended up in one category.

Conclusion

We have outlined a statistical approach to inferring lexical categories as part of a larger language learning system. We began our discussion by making reference to the need for effective tagging within a grammatical inferencing system, and further argued that because language learning is an example of a "bootstrapping" problem, lexical categories should be derived simultaneously alongside the set of rules defined over them. While the categories we obtain demonstrate a reasonably high level of similarity with traditional syntactic classes, they are inherently nonstandard and thus not intended to be evaluated in this way. It is therefore incumbent on us to offer some defense of their utility.

Cleveland [13] incorporated the lexical classification approach outlined in the previous section into a grammar induction system. He compared the results against two standard stochastic CFGs using metrics aimed at gauging the relative compactness of a grammar and its ability to place restrictions on the language it generates [14]. He found that at least one grammatical formulation based on a closed-class vocabulary produced grammars superior to the others he analysed.

It is virtually impossible to compare the lexical

category	elements			
fw_0	a	an	her	his
	my	no	one	that
	the	this	what	your
fw_1	he	I	she	then
	they	we	who	you
fw_2	are	be	had	has
	have	if	is	was
	were			
fw_3	can	could	did	do
	does	might	must	should
	will	would		
fw_4	here	him	it	me
	them	there	us	which
fw_5	all	and	as	but
	how	not	now	only
	or	than	to	when
fw_6	more	much	so	some
	very			
fw_7	about	after	at	by
	for	from	in	into
	like	of	on	out
	up	with		

Table 5: closed-class lexical categories

These examples exhibit only a weak proximity relation between an fw_0 word and the corresponding noun, because other words often intervene. However, the word positions within each noun phrase suggest that the structural roles of the words are constrained by the requirements of the phrase itself—phrases are characterized by consistent use of fw_0 words in the initial position and nouns in the final one. In order to characterize the positional roles of their constituents, a means must therefore be established to delimit phrases.

The "fw-phrase"

Determiners appear exclusively in noun phrases, and this suggests a relationship between determiner and noun [11]. Moreover, whenever determiners appear they mark the onset of a noun phrase. Consequently, since the fw_0 category can be likened to determiners, fw_0 elements can be taken to indicate the onset of some kind of phrase—a function word phrase or "fw-phrase" [12]. The phrase's left boundary is the fw_0 element itself. Generalizing, *every* closed-class word can be taken to indicate the onset of some fw-phrase type. Consequently, phrases are bounded on the right either by another closed-class word, indicating the start of a new fw-phrase, or by the end of the linguistic expression.

Three attributes define the type of a fw-phrase: the category of the closed class word that heads it, the number of words comprising it, and the category of the closed-class word that follows it.

Creating open-class categories

Every content word can be characterized by the ability it demonstrates to occupy certain structural positions in particular fw-phrase types. A structural role can be identified for each open-class word by noting the type of phrase in which it occurs and its position within that phrase. A categorial relationship can be inferred between a given open-class word and others demonstrating similar usage by analyzing the types of phrase it appears in, and which positions it occupies.

Initial open-class categories

The first stage of categorization requires that each open-class word be assigned to a temporary category. This is identified by the category of the closed-class word heading the phrase in which the open-class word appears, what follows that phrase, the length of the phrase, and the relative position occupied by the open-class word within it. For example, the sentence

A tiny bird sat in the tree

has the functional phrase structure

fw_0 tiny bird sat fw_7 fw_0 tree fw_ϕ

(where fw_ϕ marks the end of a sentence). This allows the open-class words to be assigned to temporary categories as follows:
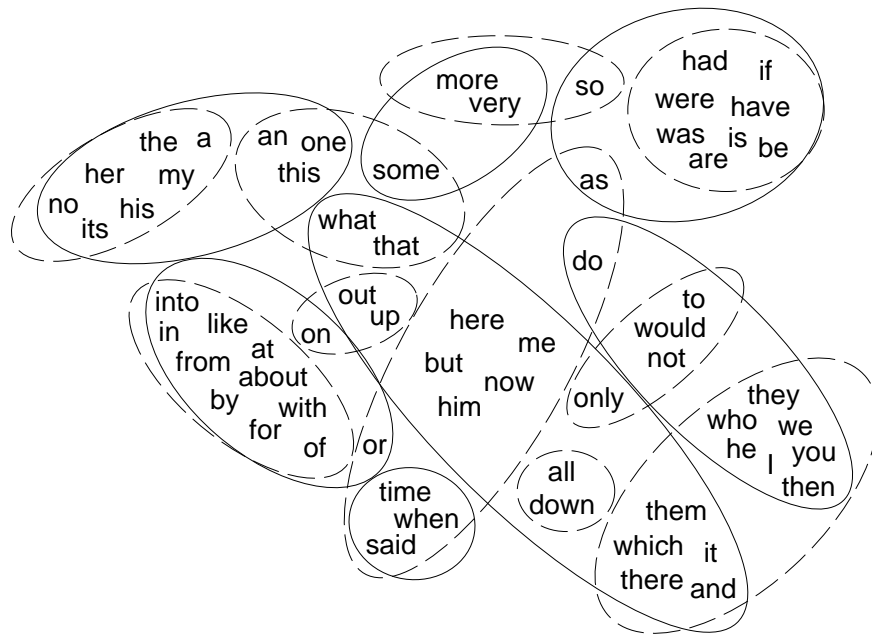
Figure 2: Clusters derived from initial random groupings

Classical categories

In contrast to the syntactically functional roles that we have supposed are fulfilled by the closed-class words, the role of open-class words is to supply content, or meaning, to text. According to classical linguistics, the categories of open-class words correspond to general types of referent. Each content lexeme conveys a particular kind of referential information, and it is the nature of its *kind* that defines the category to which the lexeme belongs.

For example, some meaning-laden words seem intuitively to function as referents to specific objects or object classes whose existence is real, surreal or imaginary. Others refer to qualities attributable to such objects—qualities like colour, texture, shape, and temperature. Still others refer to actions that can be perpetrated by or to objects. We have a nomenclature for such classical categories—nouns, adjectives, and verbs respectively—and their character is for the most part clear [11]. But our self-imposed restriction to a statistical analysis of language precludes access to any sort of semantic information that would help to assign open-class words to these categories.

Ideally, inferred lexical types should correspond closely to classical categories. Consequently, the regularities used in classical syntactic analysis prove a practical guide to the development of a suitable categorization procedure. Consider the following examples of noun usage:

> *The little brown* **fox** was quite lost.
> *An old* **man** slept on *the* **sidewalk**.
> **He** left after eating *Alison's* **lobster**.
> *Many* **people** have fed *the* **bears** from
> **car windows**.

The noun positions (in bold) demonstrate consistent occurrence as the last word of noun phrase structures (in italics). Note further that most noun phrases begin with one of the closed-class elements from the fw_0 category of Table 5.
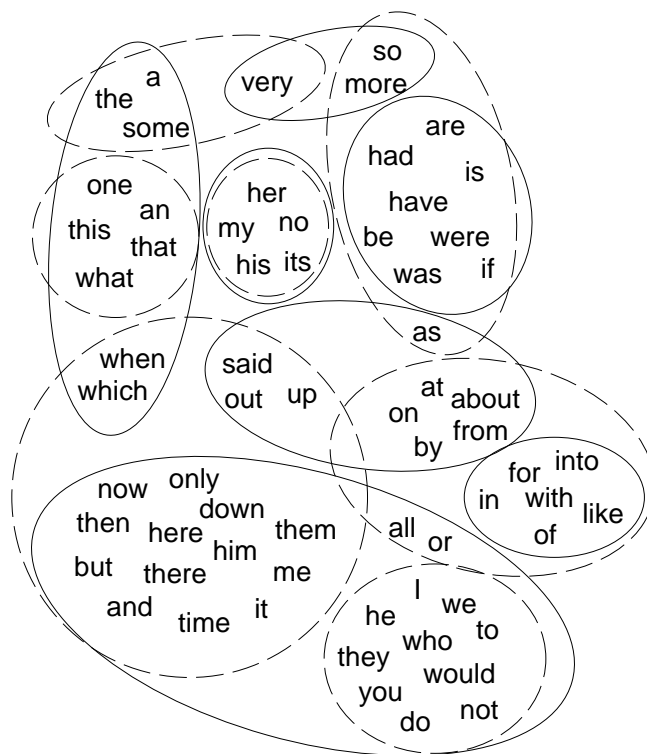
Figure 1: Categorization clusters for Hardy (solid lines) and Melville (dashed lines)

closed-class word is in its best category and, if not, reassign it. For every closed-class word, the distance is calculated to each category by averaging its first-order association probability with every word in the category. It is then reassigned to the closest category. The procedure is iterated until no reassignments occur. Figure 1 shows the final categories obtained by applying this clustering technique to the texts of Hardy and Melville. These categories do reflect functional similarities for closed-class words, particularly in the case of determiners, auxiliary verbs, prepositions, and pronouns.

Slightly different classifications are obtained depending on exactly how the procedure is carried out. For example, it is interesting to apply the iterative reassignment procedure starting from randomly-chosen initial categories. This generates the final categories shown in Figure 2.

Although rather different in detail from Figure 1, these also reflect functional similarities between closed-class words. The language inference procedures should be robust under such variation, and we believe that they are—though this has not yet been fully tested. A further categorization that was obtained is summarized in Table 5, and this is in fact the one that is used as a basis for categorizing the open-class words.

Open-class words

Every lexeme that does not qualify as closed-class is, by default, an "open-class" word. Around 99% of the vocabulary falls into this class, and it is necessary to determine a syntactic category for each of these words.

word	first-order successors	word	first-order successors	intersection size	log probability	apparent association
I	231	you	293	110	-316.0	strong
we	71	you	293	45	-238.0	strong
her	557	you	293	55	-27.7	weak
he	348	they	138	71	-253.0	strong
her	557	my	243	99	-149.0	strong
him	113	me	104	27	-149.0	strong
her	557	his	562	149	-138.0	strong
him	113	he	348	20	-18.9	weak
his	562	he	348	13	-0.1	weak
had	341	have	205	80	-211.0	strong
had	341	was	641	115	-117.0	strong
is	229	was	641	93	-117.0	strong
from	126	was	641	32	-23.1	weak
about	63	at	124	24	-184.0	strong
at	124	from	126	29	-127.0	strong
on	147	from	126	28	-101.0	strong
have	205	at	124	15	-18.9	weak
was	641	at	124	26	-15.2	weak

Table 4: Probabilities for intersection sizes (vocabulary: 11,589 words)

"you", whereas "her" and "you" are much less strongly associated. The remaining blocks of the table give samples of other associations, both strong and weak. Possessive pronouns, for example, show strong associations with each other, as do pronouns in the same case (i.e. nominative, objective, etc.). Relatively weak associations are indicated by comparisons across such class boundaries. Auxiliary verbs also show strong associations with each other, and prepositions do as well, but no cross-category relationship is indicated from the statistical evidence shown in Table 4.

Clustering closed-class words

closed-class words can be divided into syntactic categories by assuming that the strongest associations are between those whose first-order commonality is most unlikely to have arisen by chance. First, calculate the probabilities for the first-order successors' intersection sizes observed between each pair of closed-class words.

Then, place each particular word into the same syntactic category as the one to which it most strongly associated, where "strength" is measured by the unlikelihood that the two words would demonstrate such similarity in usage accidentally.

This scheme works well for most of the closed-class lexemes. However, due to a phonetic peculiarity, the words "a" and "an" exhibit a very poor first-order relationship and consequently do not end up in the same functional category. This undesirable situation could be avoided if the second-order successors could be brought into the categorization procedure, but to do this in a general way would require a scheme for weighting each of the n-order probabilities. Alternatively, if both "a" and "an" were compared with "the" before being compared with each other, they would all be categorized together. However, this would require artificial manipulation of the order of comparisons.

A third, less contrived, solution is to re-assess the initial groupings to check whether each

class word to be the set of words that immediately follow it in a particular text. (To extend the idea further, the "second-order successors" can be defined as the set of words following second after it, and so on.) The relative size of the intersection of the first-order successors of two closed-class words is a measure of how often the words are used in similar syntactic structures. Where two closed-class words share an unusually common structural usage, we assume that they are functionally similar.

To determine whether two closed-class words have a unusually large degree of commonality in their first-order successors, assume that closed-class words play no part in establishing functional roles. Then the words following each particular closed-class lexeme in a text would represent a more or less random sampling of the vocabulary.

By counting the number of different words that occur after two particular closed-class words, the expected number of different words that will appear after both can be calculated, under the assumption of random sampling. In fact, the degree of commonality is often very much higher than expected. This is no doubt partly due to the breakdown of our simplifying assumption. However, in some cases the degree of commonality—measured as the probability of this much commonality occurring by chance—is so extremely high that it indicates a substantial similarity between the syntactic roles of the two closed-class words being considered.

What is the probability that the intersection between two randomly-chosen sets is as large as a given value? Consider sets S_1 and S_2 of given sizes n_1 and n_2, whose members are drawn independently and at random from a set of size N. Denote the size of their intersection, $|S_1 \cap S_2|$, by the random variable I. It can be shown that I is distributed according to a hypergeometric distribution, and the probability that it exceeds a certain value n, $\Pr[I \geq n]$, can be determined. Unfortunately, the calculation is infeasible for large values of n_1, n_2 and N. Various approxi-

mations can be used to circumvent the problem, such as the binomial, Poisson and Normal distributions.

For example, suppose that for a particular corpus with a vocabulary of 10000 words ($N = 10000$), two particular closed-class words are both followed by 2000 different words ($n_1 = 2000$, $n_2 = 2000$). Suppose that these two sets have 700 words in common ($n = 700$). Then the Normal approximation has mean $\mu \approx 400$; in other words one expects only 400 words to be in common if the sets were randomly chosen. Its standard deviation is $\sigma \approx 16$, and so the actual figure of 700 is 19 standard deviations from the mean. It follows that the probability of I being at least as large as it is, $\Pr[I \geq 700]$, is very tiny—about 10^{-80}. (In fact tables of the Normal distribution do not generally give values for $z \geq 5$—they end with $\Pr[z > 4.99] = 3 \times 10^{-7}$.)

To estimate the probability $\Pr[I \geq n]$ in general, several approximations are possible. It was decided to split the problem into three cases depending on the size of n, n_1 and n_2. First, when $n = 0$, use $\Pr[I \geq 0] = 1$. Second, when either n_1 or n_2 is large (say n_1 or $n_2 > 100$), use the Normal approximation to the hypergeometric distribution, employing standard mathematical tables to approximate the integral that is involved. Otherwise, when both n_1 and n_2 are small (i.e. ≤ 100), calculate an approximation directly from the hypergeometric distribution and evaluate it using precomputed factorials up to 100 stored in a table.

Table 4 lists the probabilities calculated for intersection sizes of the first-order successors for some of the closed-class words in the novel *Far From the Madding Crowd*. The first line shows that "I" and "you" were followed by 231 and 293 different words respectively, of which 110 are in common. Considering the vocabulary size of 11,589 words, it is very unlikely that as many as 110 would be in common had the successors been randomly chosen—the probability is in fact only 10^{-316}! "I" and "you" thus seem to perform similar functions. So do "we" and

number of words	vocabulary items represented	fraction of vocabulary	total usage	fraction of text
1	{the}	0.01%	7,746	5.5%
2	{and, the}	0.02%	12,031	8.5%
3	{a, and, the}	0.03%	15,942	11.3%
5	{a, and, of, the, to}	0.04%	23,315	16.6%
10	{a, and, I, in, it, ...}	0.09%	32,857	23.4%
15	{a, and, as, I, in, ...}	0.13%	39,638	28.2%
115	{a, about, again, all, am, ...}	0.99%	75,688	53.8%
11589	{aaron, abandon, abasement, ...}	100.00%	140,632	100.0%

Table 2: Vocabulary distribution in *Far From the Madding Crowd*

a	for	it	or	to
about	from	its	out	up
all	had	like	said	very
an	have	me	so	was
and	he	more	some	we
are	her	my	that	were
as	him	no	the	what
at	his	not	them	when
be	I	now	then	which
but	if	of	there	who
by	in	on	they	with
do	into	one	this	would
down	is	only	time	you

Table 3: The closed class, inferred from Hardy, Melville and Carroll

Hardy, Melville, and that demonstrated in the collected works of Lewis Carroll. The resulting set is shown in Table 3, and is the one used for subsequent lexical inferencing.

Generalising the closed class

We have assumed that the relative high frequency of words ostensibly low in semanticity implies that their structural roles are functionally significant. It follows that each closed-class lexeme is either used to perform a specific and unique functional role, or is representative of one of a number of functional categories.

There are many reasons to prefer the second conclusion, even though the first permits stronger inferences. Perhaps the most compelling evidence is the intuitive notion of the functional role performed by what is called the determiner. We recognize a certain functional similarity between the words "a" and "the". In general terms, "the" is a kind of existential quantifier indicating a specific referent, whereas "a" works as a kind of universal quantifier indicating a representative of a general class of referent. Moreover, determiners like "his", "some", "many", and "all" permit reference at greater and lesser degrees of specificity.

It seems that closed-class words fall into functional categories. This is attractive because it greatly reduces the number of syntactic roles in a language. However, in keeping with a statistical analysis, we seek to achieve such generalization without relying on semantic or psychological properties.

The frequency-based method for discovering closed-class words can be regarded as a kind of zero-order test which considers the usage of words in isolation. It takes no account of the structural usage demonstrated by a word—its proximity and juxtaposition with respect to neighbors. But if closed-class words represent functional categories, then words from the same category might be expected to demonstrate similar structural usage. This can be determined by comparing the number of times each one is used in a structural context similar to that of another.

Define the "first-order successors" of a closed-

Far From The Madding Crowd vocabulary of 11589		Moby Dick vocabulary of 16832	
word	occurrences	word	occurrences
the	7746	the	13982
and	4285	of	6427
a	3911	and	6263
of	3782	a	4597
to	3591	to	4517
in	2349	in	4041
I	2123	that	2915
was	1970	his	2481
it	1566	it	2374
that	1534	I	1993
you	1468	but	1796
her	1465	he	1751
he	1391	as	1712
she	1266	with	1681
as	1191	is	1676
had	1157	was	1602
his	1145	for	1586
for	989	all	1510
with	969	this	1375
at	948	at	1297

Table 1: Most frequent words in two novels.

tain predetermined threshold. The value of the threshold is ultimately determined arbitrarily. However, we can draw on the literature to develop a rough guideline. Caplan [7] claims that "there are approximately 500 or so function words in English, and, of the 100 most common words in English, most are function words." The average person's everyday vocabulary consists of about 10,000 words. Thus the top 1% of most frequently used words from a typical vocabulary is a reasonable first approximation to the closed class—assuming that the functional importance of the other 400 words diminishes along with their declining frequency.

Table 1 provides partial lists of the most common words from the vocabularies employed by Thomas Hardy in *Far From the Madding Crowd* and Herman Melville in *Moby Dick*. A cursory analysis reveals that words used with the highest frequencies fit well with our intuitive notion of the function word. Table 2 shows that the top 1% of Hardy's vocabulary accounts for nearly 54% of the novel *Far From The Madding Crowd*, providing an indication of the extent to which such a set of words may statistically dominate an individual's vernacular.

The top 1% of any single text's vocabulary will often include some number of content (i.e. open-class) words relating to the topic of the text. For example, the word *whale* is the 28th most common word in *Moby Dick* yet it never appears in *Far From The Madding Crowd*; similarly *Bathsheba*, Hardy's 38th most frequent word, does not appear in Melville's book. In an effort to get a more representative set of statistically significant words, and thus one which might prove more syntactically useful in the general case, we define our closed class to be the intersection of the top 1% of vocabularies garnered from several disparate texts—in this case the vocabularies of

the analysis because they have no well defined syntactic category (for example, single letters of the alphabet and words connected with newsgroup administration such as *edu* and *com*).

A "divide and conquer" approach

We propose an alternative statistical method for deriving syntactic categories, simple enough to apply to the entire vocabularies of very large texts. It begins by dividing the vocabulary into two groups by some coarse method of differentiation, and thereafter continues to filter each of these groups iteratively into smaller categories according to a criterion that becomes increasingly more refined. The first distinction is made with respect to a specified level of frequency, loosely established according to linguistic notions about syntactic function. The second is made according to a statistical analysis of context with respect to the functional categories. This method, when incorporated into grammatical inference systems, has been shown to produce grammars superior to at least two standard stochastic CFGs.

Initial function-based categories

A great many languages, in particular the Indo-European languages, allow for division of their vocabulary elements into two major categories: "content" words and "function" words. Content words consist of nouns, adjectives, verbs, and so on—words whose meaning is more or less concrete and picturable. In contrast, function words are exemplified by prepositions, articles, auxiliary verbs, pronouns, and such—words whose principal role is more syntactic than semantic. Function words serve primarily to clarify relationships between the more meaning-laden elements of linguistic expression, or to introduce certain syntactic structures like verbal complements, relative clauses and questions.

Function words demonstrate many distinctive properties. Though not entirely without meaning, their semantic contribution is generally more "abstract" and less referential than that of content words. They tend not to carry stress in everyday speech. They are often the last vocabulary elements to appear in the productive language of children learning their first language. Moreover, a particular type of aphasia known as "agrammatism" is characterized by marked difficulty in the production, comprehension, and recognition of function words.

Compared to other vocabulary items, function words demonstrate high frequency usage. They tend not to enter freely into the word formation process. That is, they resist affixation and are seldom compounded with other words to form new ones. Similarly, though new content words are added to the vocabulary of a language almost daily, the number of elements in the function word class remains fixed.

The fact that the set of function words is a "closed class" of vocabulary elements that demonstrate extremely frequent usage suggests to linguists an importance in psychological processing [7, 8, 9, 10]. For lexical inference, these attributes support the idea of partitioning a vocabulary into two crude categories based on their frequency—the closed and open classes.

Closed-class words

Most linguists accept that there is a set of words that can be characterized as "closed class." But there is no consensus on exactly which words this comprises. Because lexical inference is a discovery process, membership in the closed class should be based on a criterion that identifies candidates by analysing their usage patterns. Of the previously mentioned characteristics of function words, relative high frequency is the only one that can be used to determine a closed class objectively.

Identifying the closed class

We define closed-class words operationally as those that occur more frequently than a cer-

> The cotton clothing is made of grows
> in Arkansas.

is an example of a garden-path sentence because the word "cotton" is frequently tagged as an *adjective* prior to the reader's discovery that the remainder of the sentence will not parse under this assumption. When the parse fails, the reader (presumably) correctly re-tags the word as a *noun* and is thereafter able to successfully digest the complete sentence. This particular correction may be spawned from some sort of semantic cue, but it may just as well be triggered syntactically from the lack of any parse that allows a verb to appear after the word "of" when "cotton" is tagged as an *adjective* in this sentence structure. Had "polyester" appeared in place of "grows" then, despite any objections from a semantic component, only the tagging of "cotton" as *adjective* would support a successful parse.

Static tagging

Static tagging, such as dictionary lookup, involves making arbitrary decisions about ambiguous words and often results in a large number of misclassifications. A variety of corpus-based techniques have been developed to improve tagging accuracy by incorporating contextual information [3], probabilistic modeling [4], and distributional analysis [5]—some making claims of as much as 96 or 97 percent accuracy. Unfortunately, in the case of grammar induction, an error rate of 3 or 4 percent in a tagged text may still result in an extraordinarily large number of special-case rules.

Bootstrapping

Part of the problem has been the tradition of trying to place words into standard grammatical categories like noun, verb, determiner, auxiliary verb, and so forth. Finch and Chater [6] note that the interdependence of lexical tagging and syntactic analysis indicates that grammar induction is a "bootstrapping" problem—a learning

domain where "it appears necessary to simultaneously discover a set of categories and a set of rules defined over them." That is, lexical tagging for the purpose of inferring syntactic structure should not be constrained by existing notions of syntactic categories—the rules and categories must be derived together.

Finch and Chater used the Spearman Rank Correlation Coefficient between the frequency vectors of neighbouring words to establish a hierarchical clustering for them. Using frequencies for the 10 nearest neighbours, they were able to group the 1000 most frequent words from a corpus into a relatively small number of classes using k-means clustering. By cutting the corresponding dendrogram "at a particular level of dissimilarity," they were able to produce a set of categories from which a tagged text would yield a sufficiently large number of bigrams that reliable statistics can be gathered for the inference of phrase structure.

Intractability and omission

The results obtained by Finch and Chater are encouraging, but the algorithm itself requires an extraordinary number of comparisons. The frequencies for each different word found in each of 10 nearest neighbour positions relative to each of the 1000 most frequent words must be compared with the frequencies of each different word found in each of 10 nearest neighbour positions relative to each of the other 999 most frequent words being clustered. If only one tenth of one percent of the vocabulary of a typical novel was manifest in each of the nearest neighbour sets, clustering of the 1000 most frequent words would require about 5×10^{10} comparisons. The number of comparisons required to categorize the remaining 10,000 to 20,000 words that comprise a typical novel's vocabulary would be about five orders of magnitude greater, taking a considerable amount of time to compute on even a very fast machine. Moreover, Finch and Chater indicate that a small number of items were rejected from

Probability-driven lexical classification: a corpus-based approach

Tony C. Smith

Department of Computer Science, University of Waikato, Hamilton, New Zealand
Email tcs@waikato.ac.NZ; phone: +64 (7) 838–4453; fax: +64 (7) 838–4155

Ian H. Witten

Department of Computer Science, University of Waikato, Hamilton, New Zealand
Email ihw@waikato.ac.NZ; phone: +64 (7) 838–4021; fax: +64 (7) 838–4155

keywords: language learning, syntax, lexical categories, corpus-based, statistical methods.

Abstract

Successful grammatical inference from a corpus of linguistic material rests largely on the ability to tag the words of the corpus with appropriate lexical categories. Static tagging methods, such as dictionary lookup, often misclassify words associated with multiple categories, and adaptive statistical taggers or context-based corrective taggers may still have error rates of 3 or 4 percent. Even a small proportion of lexical misclassification may lead a syntax induction mechanism to produce an extraordinarily large number of special-case rules.

By treating grammar induction as a "bootstrapping" problem in which it is necessary to simultaneously discover a set of categories and a set of rules defined over them, lexical tagging is relieved of constraints imposed by traditional notions of syntactic categories. Categories are created based on salient features in the corpus under study.

This paper describes a statistical corpus-based approach to lexical inference that uses linguistic notions of syntactic function to guide its analysis of a given text. Unlike other probability-driven inference techniques, it is simple enough to apply to the complete vocabularies of very large texts. It has been successfully incorporated into a syntactic inference system whose resulting grammars have been shown superior to two standard stochastic CFGs.

Introduction

Current research in natural language processing reveals an unmistakable trend away from the traditional AI approach of trying to develop initially comprehensive language models towards statistically driven, corpus-based language inference [1]—though linguists may see this more as a return to the quantitative techniques they used up until the mid-1950s. This trend reflects the more general move for AI as a whole away from static expert systems to the more flexible adaptive models of machine learning.

For grammar induction, this change in perspective is not quite so conspicuous, for syntactic inference is inherently adaptive—it is the gradual construction of a model that characterises regularities observed in language surface structures [2]. As it is generally held that syntax is a set of constraints on lexical categories (and larger syntactic units) rather than on the words themselves, induction is usually performed over an already tagged sequence of expressions. That is, the induction device presumes knowledge of lexical categories as part of its a priori base information. Such a presumption denies the widely held notion that sentence structure and lexical categorisation are somewhat circularly defined. For example, the following well-known sentence