Parsing, Word Associations and Typical Predicate-Argument Relations

Kenneth Church
William Gale
Patrick Hanks
Donald Hindle

Bell Laboratories and Collins

Abstract

There are a number of collocational constraints in natural languages that ought to play a more important role in natural language parsers. Thus, for example, it is hard for most parsers to take advantage of the fact that *wine* is typically *drunk, produced,* and *sold*, but (probably) not *pruned*. So too, it is hard for a parser to know which verbs go with which prepositions (e.g., *set up*) and which nouns fit together to form compound noun phrases (e.g., *computer programmer*). This chapter will attempt to show that many of these types of concerns can be addressed with syntactic methods (symbol pushing), and need not require explicit semantic interpretation. We have found that it is possible to identify many of these interesting co-occurrence relations by computing simple summary statistics over millions of words of text. This chapter will summarize a number of experiments carried out by various subsets of the authors over the last few years. The term *collocation* will be used quite broadly to include constraints on SVO (subject verb object) triples, phrasal verbs, compound noun phrases, and psycholinguistic notions of word association (e.g., *doctor/nurse*).

1. Mutual Information

Church and Hanks (1989) discussed the use of the mutual information statistic in order to identify a variety of interesting linguistic phenomena, ranging from semantic relations of the doctor/nurse type (content word/content word) to lexico-syntactic co-occurrence constraints between verbs and prepositions (content word/function word). (Jelinek (1985) has also found mutual information to be an extremely useful statistic for modeling language constraints for applications in speech recognition.) Mutual information, $I(x;y)$, compares the probability of observing word x and word y *together* (the joint probability) with the probabilities of observing x and y *independently* (chance).

$$I(x;y) \equiv \log_2 \frac{P(x,y)}{P(x)\ P(y)}$$

If there is a genuine association between x and y, then the joint probability $P(x,y)$ will be much larger than chance $P(x)\ P(y)$, and consequently $I(x;y) >> 0$, as illustrated in the table below. If there is no interesting relationship between x and y, then $P(x,y) \approx P(x)\ P(y)$, and thus, $I(x;y) \approx 0$. If x and y are in complementary distribution, then $P(x,y)$ will be much less than $P(x)\ P(y)$, forcing $I(x;y) << 0$. Word probabilities, $P(x)$ and $P(y)$, are estimated by counting the number of observations of x and y in a corpus, $f(x)$ and $f(y)$, and normalizing by N, the size of the corpus. Joint probabilities, $P(x,y)$, are estimated by counting the number of times that x is followed by y in a window of w words, $f_w(x,y)$, and normalizing by $N\ (w-1)$.[1]

1. The window size parameter allows us to look at different scales. Smaller window sizes will identify fixed expressions (idioms), noun phrases, and other relations that hold over short ranges; larger window sizes will highlight semantic concepts and other relationships that hold over larger scales.

Some Interesting Associations with ''Doctor'' in the 1987 AP Corpus (N = 15 million; w = 6)

I(x; y)	f(x, y)	f(x)	x	f(y)	y
8.0	2.4	111	honorary	621	doctor
8.0	1.6	1105	doctors	44	dentists
8.4	6.0	1105	doctors	241	nurses
7.1	1.6	1105	doctors	154	treating
6.7	1.2	275	examined	621	doctor
6.6	1.2	1105	doctors	317	treat
6.4	5.0	621	doctor	1407	bills
6.4	1.2	621	doctor	350	visits
6.3	3.8	1105	doctors	676	hospitals
6.1	1.2	241	nurses	1105	doctors
Some Less Interesting Associations with ''Doctor''					
-1.3	1.2	621	doctor	73785	with
-1.4	8.2	284690	a	1105	doctors
-1.4	2.4	84716	is	1105	doctors

2. Phrasal Verbs

Church and Hanks (1989) also used the mutual information statistic in order to identify phrasal verbs, following up a remark by Sinclair:

> ''How common are the phrasal verbs with *set*? *Set* is particularly rich in making combinations with words like *about, in, up, out, on, off,* and these words are themselves very common. How likely is *set off* to occur? Both are frequent words; [*set* occurs approximately 250 times in a million words and] *off* occurs approximately 556 times in a million words... [T]he question we are asking can be roughly rephrased as follows: how likely is *off* to occur immediately after *set*? ... This is 0.00025×0.00055 [$P(x)\ P(y)$], which gives us the tiny figure of 0.0000001375 ... The assumption behind this calculation is that the words are distributed at random in a text [at chance, in our terminology]. It is obvious to a linguist that this is not so, and a rough measure of how much *set* and *off* attract each other is to compare the probability with what actually happens... *Set off* occurs nearly 70 times in the 7.3 million word corpus [$P(x,y) = 70/(7.3\ 10^6) >> P(x)\ P(y)$]. That is enough to show its main patterning and it suggests that in currently-held corpora there will be found sufficient evidence for the description of a substantial collection of phrases... (Sinclair 1987b, pp. 151-152)

It happens that *set ... off* was found 177 times in the 1987 AP Corpus of approximately 15 million words, about the same number of occurrences per million as Sinclair found in his (mainly British) corpus. Quantitatively, $I(set;off) = 3.7$, indicating that the probability of *set ... off* is $2^{3.7} \approx 13$ times greater than chance. This association is relatively strong; the other particles that Sinclair mentions have scores of: *about* (-0.9), *in* (0.6), *up* (4.6), *out* (2.2), *on* (1.0) in the 1987 AP Corpus of 15 million words.

3. Preprocessing the Corpus with a Part of Speech Tagger

Phrasal verbs involving the preposition *to* raise an interesting problem because of the possible confusion with the infinitive marker *to*. We have found that if we first tag every word in the corpus with a part of speech using a method such as Church (1988) or DeRose (1988), and then measure associations between tagged words, we can identify interesting contrasts between verbs associated with a following preposition *to/in* and verbs associated with a following infinitive marker *to/to*. (Part of speech notation is borrowed from Francis and Kucera (1982); in = preposition; to = infinitive marker; vb = bare verb; vbg = verb + ing; vbd = verb + ed; vbz = verb + s; vbn = verb + en.) The score identifies quite a number of verbs associated in an interesting way with *to*; restricting our attention to pairs with a score of 3.0 or more, there are 768 verbs associated with the preposition *to/in* and 551 verbs with the infinitive marker *to/to*. The ten verbs found to be most associated before *to/in* are:

- *to/in*: alluding/vbg, adhere/vb, amounted/vbn, relating/vbg, amounting/vbg, revert/vb, reverted/vbn, resorting/vbg, relegated/vbn

- *to/to*: obligated/vbn, trying/vbg, compelled/vbn, enables/vbz, supposed/vbn, intends/vbz, vowing/vbg, tried/vbd, enabling/vbg, tends/vbz, tend/vb, intend/vb, tries/vbz

Thus, we see there is considerable leverage to be gained by preprocessing the corpus and manipulating the inventory of tokens.

4. Preprocessing with a Syntactic Parser

Hindle has found it useful to preprocess the input with the Fidditch parser (Hindle 1983) in order to ask about the typical arguments of verbs. Thus, for any of verb in the sample, we can ask what nouns it takes as subjects and objects. The following table shows the objects of the verb *drink* that appeared at least two times in a sample of six million words of AP text, in effect giving the answer to the question ‘‘what can you drink?’’ Calculating the co-occurrence weight for *drink*, shown in the third column, gives us a reasonable ranking of terms, with *it* near the bottom. This list of drinkable things is intuitively quite good.

Object	Frequency	Mutual Information
<quantity> beer	2	12.34
tea	4	11.75
Pepsi	2	11.75
champagne	4	11.75
liquid	2	10.53
beer	5	10.20
wine	2	9.34
water	7	7.65
anything	3	5.15
much	3	2.54
it	3	1.25
<quantity>	2	1.22

A standard alternative approach to the classification of entities is in terms of a hierarchy of types. The biological taxonomy is the canonical example: a penguin is a bird is a vertebrate and so on. Such ‘‘is-a’’ hierarchies have found a prominent place in natural language processing and knowledge representation because they allow generalized representation of semantic features and of rules. There is a wide range of problems and issues in using ‘‘is-a’’ hierarchies in natural language processing, but two especially recommend that we investigate alternative classification schemes like the one reported here. First, ‘‘is-a’’ hierarchies are large and complicated and expensive to acquire by hand. Attempts to automatically derive these hierarchies for words from existing dictionaries have been only partially successful (Chodorow, Byrd, and Heidorn 1985). Yet without a comprehensive hierarchy, it is difficult to use such classifications in the processing of unrestricted text. Secondly, for many purposes, even knowing the subclass-superclass relations is insufficient; it is difficult to predict which properties are inherited from a superclass and which aren’t, and what properties are relevant in a particular linguistic usage. So for example, as noted above, despite the fact that both potatoes and peanuts are edible foods that grow underground, we typically *bake potatoes*, but *roast peanuts*. A distribution-based classification, if successful, promises to do better at least on these two problems.

5. Significance Levels

If the frequency counts are very small, the mutual information statistic becomes unstable. This is the reason for not reporting objects that appeared only once with the verb *drink*. Although these objects have very large mutual information scores, there is also a very large chance that they resulted from some quirk in the corpus, or a bug in the parser. For some purposes, it is desirable to measure confidence rather than likelihood. Gale and Church have investigated the use of a t-score instead of the mutual information score, as a way of identifying ‘‘significant’’ bigrams.

The following table shows a few significant bigrams ending with *potatoes*, computed from 44 million words of AP news wire from 2/12/88 until 12/31/88. The numbers in the first column indicate the confidence in standard deviations that the word sequence is interesting, and cannot be attributed to chance.

t	*x*	*y*
4.6	sweet	potatoes
4.3	mashed	potatoes
4.3	,	potatoes
4.0	and	potatoes
3.8	couch	potatoes
3.3	of	potatoes
3.3	frozen	potatoes
2.8	fresh	potatoes
2.8	small	potatoes
2.1	baked	potatoes

These numbers were computed by the following formula

$$t = \frac{E(Pr(x\ y)) - E(Pr(x)\ Pr(y))}{\sqrt{\sigma^2(Pr(x\ y)) + \sigma^2(Pr(x)\ Pr(y))}}$$

where $E(Pr(x\ y))$ and $\sigma^2(Pr(x\ y))$ are the mean and variance of the probability of seeing word x followed by word y. The means and variances are computed by the Good-Turing method (Good 1953).

Let r be the number of times that the bigram $x\ y$ was found in a corpus of N words, and let N_r be the frequencies of frequences (the number of bigrams with count r). Then r^*, the estimated expected value of r in similar corpus of the same size, is $r^* = N \times E(Pr(x\ y)) = (r+1)\ \frac{N_{r+1}}{N_r}$ and the variance of r is $\sigma^2(r) = N^2\sigma^2(Pr(x\ y)) = r^*\ (1 + (r+1)^* - r^*)$

6. Just a Powerful Tool

Although it is clear that the statistics discussed above can be extremely powerful aids to a lexicographer, they should not be overrated. We do not aim to replace lexicographers with self-organizing statistics; unlike (Jelinek, 1985), we merely hope to provide a set of tools that could greatly improve their productivity. Suppose, for example, that a lexicographer wanted to find a set of words that take sentential complements. Then it might be helpful to start with a table of t-scores such as:

t	x	y
74.0	said	that
50.9	noted	that
43.3	fact	that
41.9	believe	that
40.7	found	that
40.1	is	that
40.0	reported	that
39.5	adding	that
38.6	Tuesday	that
38.4	Wednesday	that

It might be much quicker for a lexicographer to edit down this list than to construct the list from intuition alone. It doesn't take very much time to decide that *Tuesday* and *Wednesday* are less interesting than the others. Of course, it might be possible to automate some of these decisions by appropriately preprocessing the corpus with a part of speech tagger or a parser, but it will probably always be necessary to exercise some editorial judgment.

7. Practical Applications

The proposed statistical description has a large number of potentially important applications, including:

- enhancing the productivity of lexicographers in identifying normal and conventional usage,
- enhancing the productivity of computational linguists in compiling lexicons of lexico-syntactic facts,
- providing disambiguation cues for parsing highly ambiguous syntactic structures such as noun compounds, conjunctions, and prepositional phrases,
- retrieving texts from large databases (e.g., newspapers, patents), and
- constraining the language model both for speech recognition and optical character recognition (OCR).

Consider the optical character recognizer (OCR) application. Suppose that we have an OCR device such as (Kahan, Pavlidis, Baird 1987), and it has assigned about equal probability to having recognized ‘‘farm’’ and ‘‘form,’’ where the context is either: (1) ‘‘federal ___ credit’’ or (2) ‘‘some ___ of.’’ We doubt that the reader has any trouble specifying which alternative is more likely. By using the following probabilities for the eight bigrams in this sequence, a computer program can rely on an estimated likelihood to make the same distinction.

x	*y*	*Observations per million words*
federal	farm	0.50
federal	form	0.039
farm	credit	0.13
form	credit	0.026
some	form	4.1
some	farm	0.63
form	of	34.0
farm	of	0.81

The probability of the tri-grams can be approximated by multiplying the probabilities of the the two constituent bigrams. Thus, the probability of *federal farm credit* can be approximated as $(0.5\times10^{-6})\times(0.13\times10^{-6}) = 0.065\times10^{-12}$. Similarly, the probability for *federal form credit* can be approximated as $(0.039\times10^{-6})\times(0.026\times10^{-6}) = 0.0010\times10^{-12}$. The ratio of these likelihoods shows that ‘‘farm’’ is $(0.065\times10^{-12})/(0.0010\times10^{-12}) = 65$ times more likely than ‘‘form’’ in this context. In the other context, ‘‘some ___ of,’’ it turns out that ‘‘form’’ is 273 times more likely than ‘‘farm.’’ This example shows how likelihood ratios can be used in an optical character recognition system to disambiguate among optically confusable words. Note that alternative disambiguation methods based on syntactic constraints such as part of speech are unlikely to help in this case since both ‘‘form’’ and ‘‘farm’’ are commonly used as nouns. In contrast, the tri-gram model (Jelinek, 1985) will have no difficulty with this example. Although the tri-gram model may be obviously flawed (Chomsky, 1957, p. 17), it does have its strengths.

8. Alternatives to Collocation for Recognition Applications

There have been quite a number of attempts to use syntactic methods in speech recognition, beginning with the ARPA speech project and continuing on to the present. It might be noted, however, that there has not been very much success, perhaps because syntax alone is not a strong enough constraint on language use (performance). We believe that collocational constraints should play an important role in recognition applications, and attempts to ignore collocational constraints and use purely syntactic methods will probably run into difficulties.

Syntactic constraints, by themselves, though are probably not very important. Any psycholinguist knows

that the influence of syntax on lexical retrieval is so subtle that you have to control very carefully for all the factors that really matter (e.g., word frequency, word association norms, etc.). On the other hand, collocational factors (word associations) dominate syntactic ones so much that you can easily measure the influence of word frequency and word association norms on lexical retrieval without careful controls for syntax.

There are many ways to demonstrate the relative lack of constraint imposed by syntax. Recall the old television game show, ''The Match Game,'' where a team of players was given a sentence with a missing word, e.g, ''Byzantine icons could murder the divine BLANK,'' and asked to fill in the blank the same way that the studio audience did. The game was 'interesting' because there are enough constraints in natural language so that there is a reasonably large probability of a match. Suppose, however, that we make our speech recognition device play the match game with a handicap; instead of giving the speech recognition device the word string, ''Byzantine icons could murder the divine BLANK,'' we give the speech recognition device just the syntactic parse tree, [S [NP nn nns] [VP [AUX md] v [NP at jj BLANK]]], and ask it to guess the missing word. This is effectively what we are doing by limiting the language model to syntactic considerations alone. Of course, with this the handicap, the match game isn't much of a game; the recognition device doesn't have a fair chance to guess the missing word.

We believe that syntax will ultimately be a very important source of constraint, but in a more indirect way. As we have been suggesting, the real constraints will come from word frequencies and collocational constraints, but these questions will probably need to be broken out by syntactic context. How likely is it for this noun to conjoin with that noun? Is this noun a typical subject of that verb? And so on. In this way, syntax plays a crucial role in providing the relevant representation for expressing these very important constraints, but crucially, it does not provide very much useful constraint (in the information theoretic sense) all by itself.[2]

9. Concluding Remarks

In any natural language there are restrictions on what words can appear together in the same construction, and in particular, on what can be arguments of what predicates. It is common practice in linguistics to classify words not only on the basis of their meanings but also on the basis of their co-occurrence with other words. Running through the whole Firthian tradition, for example, is the theme that ''You shall know a word by the company it keeps'' (Firth, 1957).

> ''On the one hand, *bank* co-occurs with words and expressions such as *money, notes, loan, account, investment, clerk, official, manager, robbery, vaults, working in a, its actions, First National, of England*, and so forth. On the other hand, we find *bank* co-occurring with *river, swim, boat, east* (and of course *West* and *South*, which have acquired special meanings of their own), *on top of the*, and *of the Rhine*.'' (Hanks 1987, p. 127)

Harris (1968) makes this ''distributional hypothesis'' central to his linguistic theory. His claim is that: ''the meaning of entities, and the meaning of grammatical relations among them, is related to the restriction of combinations of these entities relative to other entities,'' (Harris 1968:12). Granting that there must be some relationship between distribution and meaning, the exact nature of such a relationship to our received notions of meaning is nevertheless not without its complications. For example, there are some purely collocational restrictions in English that seem to enforce no semantic distinction. Thus, one can *roast chicken* and *peanuts* in an oven, but typically *fish* and *beans* are *baked* rather than *roasted*: this fact seems to be a quirk of the history of English. Polysemy provides a second kind of complication. A *sentence* can be *parsed* and a *sentence* can be *commuted*, but these are two distinct senses of the word *sentence*; we

2. Much of the work on language modeling for speech recognition has tended to concentrate on search questions. Should we still be using Bates' island driving approach (Bates 1975), or should we try something newer such as Tomita's so-called generalized LR(k) parser (Tomita 1986)? We suggest that the discussion should concentrate more on describing the facts, and less on how they are enforced.

should not be misled into positing a class of things that can be both *parsed* and *commuted.*

Given these complicating factors, it is by no means obvious that the distribution of words will directly provide a useful semantic classification, at least in the absence of considerable human intervention. The work that has been done based on Harris' distributional hypothesis (most notably, the work of the associates of the Linguistic String Project (see for example, Hirschman, Grishman, and Sager 1975)) unfortunately does not provide a direct answer, since the corpora used have been small (tens of thousands of words rather than millions) and the analysis has typically involved considerable intervention by the researchers. However, with much larger corpora (10-100 million words) and robust parsers and taggers, the early results reported here and elsewhere appear extremely promising.

Acknowledgements

The first author has benefited greatly from a 1982 course taught at MIT by Jelinek, Bahl and Mercer.

References

Bates, M., ''Syntactic Analysis in a Speech Understanding System,'' BBN Report No. 3116, 1975.

Chodorow, M, Byrd, R., and Heidorn, G., (1985) ''Extracting semantic hierarchies from a large on-line dictionary,'' ACL Proceedings.

Chomsky, N., (1957), ''Syntactic Structures,'' The Hague: Mouton & Co.

Church, K., (1988), ''A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text,'' Second Conference on Applied Natural Language Processing, Austin, Texas.

Church, K., and Hanks, P., (1989), ''Word Association Norms, Mutual Information, and Lexicography,'' ACL Proceedings.

DeRose, S., ''Grammatical Category Disambiguation by Statistical Optimization,'' Computational Linguistics, Vol. 14, No. 1, 1988.

Firth, J., (1957), ''A Synopsis of Linguistic Theory 1930-1955'' in *Studies in Linguistic Analysis*, Philological Society, Oxford; reprinted in Palmer, F., (ed. 1968), *Selected Papers of J.R. Firth*, Longman, Harlow.

Francis, W., and Kucera, H., (1982), *Frequency Analysis of English Usage,* Houghton Mifflin Company, Boston.

Good, I. J., (1953), *The Population Frequencies of Species and the Estimation of Population Parameters*, Biometrika, Vol. 40, pp. 237-264.

Hanks, P., (1987), ''Definitions and Explanations,'' in Sinclair (1987a).

Harris, Z., (1968), ''Mathematical Structures of Language,'' New York: Wiley.

Hirschman, L., Grishman, R., and Sager, N., (1975) ''Grammatically-based automatic word class formation,'' *Information Processing and Management*, 11, 39-57.

Hindle, D., (1983), ''User manual for Fidditch, a deterministic parser,'' Naval Research Laboratory Technical Memorandum #7590-142

Jelinek, F., (1985), ''Self-organized Language Modeling for Speech Recognition,'' IBM Report.

Kahan, S., Pavlidis, T., and Baird, H., (1987) ''On the Recognition of Printed Characters of any Font or Size,'' IEEE Transactions PAMI, pp. 274-287.

Sinclair, J., Hanks, P., Fox, G., Moon, R., Stock, P. (eds), (1987a), *Collins Cobuild English Language Dictionary*, Collins, London and Glasgow.

Sinclair, J., (1987b), ‘‘The Nature of the Evidence,’’ in Sinclair, J. (ed.), *Looking Up: an account of the COBUILD Project in lexical computing*, Collins, London and Glasgow.

Tomita, M., (1986), *Efficient Parsing for Natural Language*, Kluwer Academic Press.