Tagging for Learning: Collecting Thematic Relations from Corpus

Uri Zernik and Paul Jacobs
Artificial Intelligence Program
GE Research and Development Center
Schenectady, NY 12301
USA

Abstract

Recent work in text analysis has suggested that data on words that frequently occur together reveal important information about text content. Co-occurrence relations can serve two main purposes in language processing. First, the statistics of co-occurrence have been shown to produce accurate results in syntactic analysis. Second, the way that words appear together can help in assigning thematic roles in semantic interpretation. This paper discusses a method for collecting co-occurrence data, acquiring lexical relations from the data, and applying these relations to semantic analysis.

1 Introduction

Two text processing problems rely heavily on co-occurrence patterns– the way that words appear together, possibly idiosyncraticly. First, statistically weighted co-occurrence information can assist in the "bracketing" of noun groups, which can otherwise lead to a combinatoric explosion of parse trees [1]. Second, co-occurrence relations can provide evidence of semantic information for thematic-role assignment, an important task that is otherwise fraught with inaccuracy.

Only co-occurrence patterns collected over a corpus can help to determine which is *object* and which is *recipient* in PAID DIVIDEND (IS SECURE) vs. PAID SHAREHOLDERS (ARE SATISFIED). A sufficiently rich lexicon would include the semantic preferences for distinguishing these thematic roles, but such a lexicon does not yet exist.

Co-occurrence patterns are a means of probing a global corpus for clues that help resolve ambiguity at the local sentence level. Patterns such as PAID TO SHAREHOLDERS and PAID THEM THE DIVIDEND are detected in the corpus at large. Through these latter examples, in which the distinction between *recipient* and *object* relative to the *dative* verb PAY is made explicit, the former cases in which the relation is implicit can be resolved.

In contrast to previous work which addressed the identification of surface relations, i.e., SVO triples [2], in our work we address the acquisition of semantic relations, focussing at the assigment of thematic roles. This task (i.e. tagging for acquisition) requires high reliability and so it relies less on statistical properties and more on deterministic local marking.

In this paper we discuss a technique for parsing and semanticly analyzing complex sentences with the aid of co-occurrence relations, and show how these relations are acquired from tagged corpus.

1.1 The Phenomenon

Consider, for example, the sentence below, taken from the Dow-Jones newswire:

```
THE LARGEST COMPANY ON THE LIST,
WHICH LAST PAID SHAREHOLDERS IN JANUARY,
SAID THE 5 PC STOCK  DIVIDEND WOULD BE
PAYABLE FOLLOWING THE PAYMENT OF THE
CASH DIVIDEND.  (DJ, October 27, 1988)
```

For this sentence, which is not exotic or unusual in its complexity, there are 24 non-trivial different parse trees. Human readers, in contrast to most programs, can quickly identify groups of words that "hang together" such as COMPANY PAID A DIVIDEND, STOCK DIVIDEND, and CASH DIVIDEND, and use these clusters to understand the sentence unambiguously. Moreover, a human reader can easily recognize SHAREHOLDERS as recipient and DIVIDEND as the object of PAY. Along these lines, our program develops the capability to identify such patterns by training on a large corpus of examples.

1.2 The Training Corpus

The training corpus, from which our lexical information is extracted, consists of more than ten mil-

lion words from the Dow Jones newswire (10 months worth of stories). For the root PAY, for instance, we collected more than 6000 examples, 20 of which are given below.

To exploit this data, a system must transform common patterns into operational templates, encoding a core relation between the words. The sections that follow describe the evolution and implementation of this acquisition technique.

2 Co-occurrence: Previous Work

Garside [4] and Church et al. [1] provided a major impetus for this line of work. In Church's work, a collection of English collocations bootstrapped from a tagged corpus facilitated the construction of an adaptive "tagger", a program that annotates a text with part-of-speech information.

Frank Smadja [7] continued Church's effort by collecting *operational pairs* such as verb-noun and adjective-noun pairs. Smadja used these pairs to constrain lexical choice in a language generator; for example, the system prefers "*deposit* a check" to "*place* a check" based on the frequency of co-occurrence of *deposit* and *check*.

Ido Dagan [3] pursued this topic further by projecting co-occurrences beyond the local context, using collocations for anaphora resolution. For example in,

```
THE CAR WAS DRIVING ON THE ROAD.
   SUDDENLY IT BRAKED.
```

CAR is selected over ROAD as the anaphor of IT, since CAR BRAKE is a stronger collocation than ROAD BRAKE. Interestingly, this idea complements Wilks' *preference semantics* [8], in which preference is based on a semantic hierarchy. In Dagan's method, preferences are based on word patterns acquired from corpus.

Our work further emphasizes global-sentence connections. An example that highlights the use of co-occurrence is given on the next page.

```
THE CHAIRMAN AND CHIEF EXECUTIVE OF FRANKL-
IN FIRST FEDERAL SAVINGS & LOAN ASSOCIAT-
ION OF WILKES-BARRE, [SAID] FRANKLIN FIRST
FEDERAL'S PLAN OF CONVERSION HAD BEEN
APPROVED BY THE FEDERAL HOME LOAN BANK
BOARD [AND THAT] THE OFFERING OF COMMON
SHARES IN FRANKLIN FIRST FINANCIAL CORP.
HAD BEEN APPROVED BY THE BANK BOARD AND BY
THE SEC. (DJ, 07-25-88).
```

What is the attachment of THAT? THAT could potentially attach to almost any preceding word, e.g., FEDERAL THAT, BOARD THAT, CONVERSION THAT, SAID THAT, etc. The affinity of the word pair SAY THAT (although it does not appear in this sentence as a collocation) supports the appropriate attachment.

Furthermore, co-occurrence relations support thematic-role assignment. This is important for our ultimate objective of producing more accurate conceptual information from news stories [5]. The text below illustrates one type of problem in role assignment:

```
   THE LARGEST COMPANY ON THE LIST,
WHICH LAST PAID SHAREHOLDERS IN JANUARY,
SAID THE 5 PC STOCK DIVIDEND WOULD BE
PAYABLE FOLLOWING THE PAYMENT OF THE
CASH DIVIDEND. (DJ, October 27, 1988)
```

Who paid what to whom and when? Co-occurrence-based analysis generates lexical relations such as *subj-verb*, *verb-obj*, and *verb-obj2*, relations which are further mapped into appropriate thematic and semantic roles. The program thus determines that COMPANY is the payer of PAID, SHAREHOLDERS the payee, and DIVIDEND the payment.

3 Lexical Representation

An acquired lexical structure called a *Thematic Relations* (Figure 2) facilitates this analysis. For a pair of content words, a relation provides (1) a strength of association (or "mutual affinity"), and (2) a structure type.

This table is acquired from corpus by a tagger based on morphology and local syntax.

4 Extracting Co-occurrence Relations from Corpus

The algorithm operates in three steps: (1) tag the corpus for morphology and part of speech, (2) collect collocations using relative frequency, and (3) use tagging to determine lexical relations within collocations.

4.1 Part-of-speech Tagging

Since the corpus size is about 10-million words, a full-fledged global sentence parsing is prohibitively expensive, and tagging must be carried out by *localist* methods, i.e., by means of morphology and local syntactic markers. There are three degrees of difficulty of cases to be tagged.

Morphology-based Tagging: Only a few words can be tagged using morphology alone. While PAYMENT and SHAREHOLDERS are unambiguously nouns, morphology-based tagging is ambiguous for most words. For example, PAID and SAID could be either verb or adjective (i.e. participle modifier); STOCK and CASH could be either noun or verb.

REEMENT, IT HAS AGREED NOT TO PAY ANY FUTURE CASH DIVIDENDS, INCLUDING THE
D THAT IT INTENDS TO CONTINUE PAYING THE DIVIDEND. -0-; 11 08 AM EDT 07-22-
TIONS AND MODIFIYING DIVIDEND PAYING A STOCK OF 60 CENTS FOR A TOTAL OF $1.
A PATTERN FOR THE FUTURE. IT PAID A SPECIAL DIVIDEND OF 8C LAST YEAR. -0-
JUNE 30. THE COMPANY LAST PAID A 7.5C DIVIDEND ON MAY 9. GROW GROUP
A 10 PC STOCK DIVIDEND TO BE PAID AUG. 15. -0-; 2 09 PM EDT 07-28-88:"?
N INCOME DIVIDEND OF 1C A SHR PAID IN FEBRUARY. -0-; 3 10 PM EDT 07-28-88:
AUG. 15. THE COMPANY LAST PAID A 10C SPECIAL DIVIDEND IN SEPTEMBER 1987
UT THE SPECIAL DIVIDEND TO BE PAID FROM PROCEEDS OF THE SALE TO $6 A SHARE
CT. 21. THE COMPANY LAST PAID A DIVIDEND OF 11 CENTS A SHARE ON JULY 2
10 PER SHARE SPECIAL DIVIDEND PAID TO STOCKHOLDERS ON JAN. 5, 1988. TOPP
PER SHARE. THE DIVIDEND IS PAYABLE TO SHAREHOLDERS OF RECORD JULY 5.
TED FOR A 5 PC STOCK DIVIDEND PAYABLE AUG. 12 TO HOLDERS OF RECORD JULY 15.
ERLY DIVIDEND OF 68.75 CENTS PAYABLE OCT. 1 WILL BE PAYED IN THE USUAL MAN
TERLY DIVIDEND OF 12 CENTS IS PAYABLE AUG. 29 TO HOLDERS OF RECORD AUG. 12.
HE SPLIT AND THE DIVIDEND ARE PAYABLE SEPT. 14 TO HOLDERS OF RECORD AUG. 22
1.5 MILLION. THE DIVIDEND IS PAYABLE AUG. 18 TO HOLDERS OF RECORD AUG. 8.
F THE COMPANY ON ANY DIVIDEND PAYMENT DATE ON OR AFTER AUG. 1, 1990, FOR TH
N THE UPCOMING FINAL DIVIDEND PAYMENT OF 10.85 PENCE A SHARE. HEIGHTENING
LDING ONE ADDITIONAL DIVIDEND PAYMENT OVER A 12-MONTH PERIOD. DUE THURSDAY.

Figure 1: PAY Sentences in Corpus

0.15	predicate:PAY	subject:COMPANY
0.56	predicate:PAY	object:DIVIDEND
0.73	predicate:PAY	object2:SHAREHOLDER
0.11	predicate:PAY	object:MILLION
0.19	predicate:PAY	object:CASH
0.22	predicate:PAY	object:*number* PC
0.46	predicate:PAY	object:TOP RATES

Figure 2: Word Pairs Indicating Mutual Affinity and Thematic Roles

Syntax-based Tagging: Local syntactic markers help to remove most cases of ambiguity. For example, was SAID (read: the word SAID preceded by was) can be unambiguously tagged a verb; the PAID shareholders, is an adjective; and the STOCK is definitely a noun.

Statistics-Based Tagging: Taggers reported by [4; 1] have capitalized on a large collection of bigrams plus statistically weighted grammar rules. In this method, statistical properties are acquired from a large training corpus which was tagged manually. Statistical methods have proved very effective, and attained a high level of accuracy [6].

4.2 Problematic Cases

Some cases prove even more difficult and cannot be resolved by localist methods. Consider the following two examples.

- "The company preferred stock PAID ..." . In this clause, PAID, could be either an adjective or a verb (see "the horse raced past the barn"). Indeed, this clause could probably be determined by a global parse, however, this would be too expensive computationally.
- "CONVINCING MANAGEMENT proved tough" is even harder since it presents a Necker cube situation (i. e. changing the interpretation of either word seems immediately to change the interpretation of the pair). Is it an adjective-noun or is it a verb-noun pair? In general, the analysis of such pairs requires deeper understanding of word relationships. Consider another example:

  LATER IN THE DAY BUYING INTEREST DIMINISHED ...

  Again, it is difficult to tell whether INTEREST in BUYING diminished or the BUYING of INTERESTs diminished. Thus, local clues do not contribute towards the proper resolution of such cases.

The incorrect resolution of such cases, which unfortunately are pervasive in the corpus, impinges on two objectives: performance and learning.

In order to perform text analysis, in the first case one must determine whether management was convinced, or the management convinced some second party; in the second case, one must determine the subject of the main verb of the sentence, i.e., which is the subject of DIMINISHED? Many applications require an unambiguous result. Thus a call must be made one way or another. Statistical means might make that call slightly more judiciuos on the average.

However, when tagging is used for learning of thematic roles, inappropriate resolution of such cases can drastically contaminate the final results by biasing it in a certain direction. Results are far more accurate when ambiguous cases are left out altogether.

4.3 Tagging for Learning

Our tagger is based on a 7,000-root lexicon that facilitates accurate morphological analysis, and about 100 local-syntax rules. It produces tagging for about 60% of the content words in the corpus. Tagged output for a sample sentence is given below.

```
THE///DT LARGEST/LARGE/EST/AD COMPANY///NN
ON///PP THE///DT LIST///NN *comma*///SP
WHICH///CC LAST///AD PAID/PAY/ED/? SHAREHO-
LDERS/SHAREHOLDER/S/NN IN///PP JANUARY///DD
*comma*///CC SAID/SAY/ED/?? THE///DT 5///AD
PC///NN STOCK///?? DIVIDEND///NN WOULD/WILL
//AX BE///AX PAYABLE/PAY/ABLE/AD FOLLOWING/
FOLLOW/ING/?? THE///DT PAYMENT/PAY/MENT/NN
OF///PP THE///DT CASH///NN DIVIDEND///NN
*period*///SP
```

A 4-tuple in the sentence above is a word/root/affix/part-of-speech. As expected, many content words in this sentence cannot be unambiguously tagged, and are marked ?, i.e., undetermined. In particular, notice that PAID remains unresolved.

Fortunately, most PAY cases in the corpus are simpler and are appropriately tagged.

```
OF///PP THE///DT CASH///NN DIVIDEND///NN
THE///DT COMPANY///NN LAST///JJ PAID/PAY/ED
/VA A///DT 5///NN DIVIDEND///NN ON///PP JA-
NUARY///DD ...
```

For purposes of thematic role acquisition the identification of passive and active voice is crucial. In the sample sentence above, PAID is appropriately tagged as a verb in the active voice (marked as **VA**).

4.4 Collecting Collocations

Based on the tagging above (the root field), all collocations in the corpus are counted, and the following table is generated.

This table is similar to Smadja's [7], and it provides the position of collocative words relative to PAY, and the total count within 4 words in either direction.

4.5 Determining Lexical Relations

Lexical relations are determined using the known functionality of the verb (see [9]) and supporting examples. PAY is marked in the lexicon as a dative verb.

Consider 5 cases containing the pair PAY SHAREHOLDER, from which the thematic relation is induced (**VA** stands for verb, active voice; **VP** for verb, passive voice; **AD** for adjective).

word	-4	-3	-2	-1	0	+1	+2	+3	+4	total
PRICE	5	14	438	38	0	17	12	32	12	558
COMPANY	47	53	71	26	0	2	6	1	161	367
DIVIDEND	37	42	36	121	0	11	1	14	25	287
RATE	6	5	16	109	0	14	112	16	3	281
MILLION	9	28	12	2	0	4	102	53	53	263
STOCK	35	0	134	2	0	7	1	22	2	203
MAJOR	0	2	0	6	0	2	0	92	80	182
DUE	1	4	35	16	0	4	39	66	7	172
INTEREST	1	3	5	74	0	8	14	29	34	168
SPECIAL	13	5	5	84	0	3	17	9	24	160
CASH	3	11	9	71	0	3	8	23	17	145
CENT	19	26	10	11	0	3	33	26	10	138
SHARE	9	25	0	29	0	4	7	23	33	130
AMOUNT	24	43	15	10	0	3	1	18	16	130
PC	12	30	14	23	0	4	2	21	11	117
SPLIT	2	10	25	57	0	0	4	0	0	98
DATE	29	0	1	3	0	22	29	10	1	95

Figure 3: A Distance Matrix between Word Pairs

```
(1)   STINGHOUSE SAID IT INTENDS TO PAY/va THE TWO SHAREHOLDERS/nn $2.08 A SHARE PLUS A
(2)   ONTROL OF THE COMPANY WITHOUT PAYING/va ALL SHAREHOLDERS/nn A FAIR PRICE.    THE
(3)   THE CASH PORTION OF THE PRICE PAID/?? TO POLYSAR COMMON SHAREHOLDERS/nn WILL INCR
(4)   CIPATING SHAREHOLDERS/nn WILL BE PAID/vp $3 A SHARE CASH.  NO BROKERAGE FEES OR T
(5)   PER SHARE.    THE DIVIDEND IS PAYABLE/ad TO SHAREHOLDERS/nn OF RECORD JULY 5.
(6)   ENTS A SHARE FROM 37.5 CENTS, PAYABLE/ad SEPT. 1 TO SHAREHOLDERS/nn OF RECORD AUG
```

Figure 4: Word Pairs Tagged as to their Part of Speech

Examples (1), (4), and (5) support the hypothesis that SHAREHOLDER is an object2 (the recipient) of PAY.

5 Current Status and Conclusions

Based on a number of tagged sentences, the system determines that SHAREHOLDERS are recipients of PAY, while DIVIDENDS are objects. This generalized lexical relation enables the semantic resolution of more difficult cases such as DIVIDEND PAYMENT and COMPANY PAID STOCK DIVIDEND.

The implemented system using these techniques includes several elements: (1) morphology analysis - currently produces accurate results for all the required cases; (2) tagging - produces results for only 60% of the required examples; more detailed rules could improve this figure to about 70%; (3) rule forming - currently works only with dative verbs such as PAY and SELL.

A number of important pieces of recent research have highlighted the power of co-occurrence information in text. In the techniques described here, we have extended this research to use co-occurrence information for discriminating thematic roles. These techniques combine data acquisition from a tagged corpus with relation-driven language analysis to derive thematic knowledge from the text.

References

[1] K. Church, W. Gale, P. Hanks, and D. Hindle. Parsing, word associations, and predicate-argument relations. In *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, 1989.

[2] K. Church, W. Gale, P. Hanks, and D. Hindle. Using statistics in lexical analysis. In U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

[3] I. Dagan. Using collocation in anaphora resolution. Technical report, Technion, Computer Science Deptartment, Haifa, Israel, 1989.

[4] G. Leech R. Garside and G. Sampson. *The Computational Analysis of English*. Longman, London, Britain, 1987.

[5] Lisa F. Rau, Paul S. Jacobs, and Uri Zernik. Information extraction and text summarization using linguistic knowledge acquisition. *Information Processing and Management*, 25(4):419–428, 1989.

[6] B. Santorini. Annotation manual for the pen treebank project. Technical report, University of Pennsylvania, Computer and Information Science, Philadelphia, PA, 1990.

[7] F. Smadja. Macrocoding the lexicon with co-occurrence knowledge. In U. Zernik, editor, *First International Lexical Acquisition Workshop*. 1989.

[8] Y. Wilks. A preferential, pattern-matching semantics for natural language understanding. *Artificial Intelligence*, 6, 1975.

[9] U. Zernik. Lexical acquisition: Learning from corpus by capitalizing on lexical categories. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, Michigan, 1989.