BROWN UNIVERSITY
Department of Computer Science
Master's Project

CS-96-M4

"Extracting Grammatical Information From Large Corpora

by

Murat Ersan

Extracting Grammatical Information From Large Corpora

Murat Ersan

Department of Computer Science
Brown University

Submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Computer Science at Brown University

August 1995

Professor Eugene Charniak
Advisor

Extracting Grammatical Information From Large Corpora

Murat Ersan and Eugene Charniak
Department of Computer Science, Brown University
Providence, RI 02906

April 23, 1995

Abstract

In this paper we tried to describe some implemented programs that extract grammatical information about English from a large untagged corpus. Using statistical techniques we generated the case frames of verbs, and the obligatory and optional prepositions attached to nouns and adjectives, which we believe are of great importance for parsers and text-generators.

1 Introduction

The main objective of this project is to extract grammatical information from large corpora using statistical techniques. We mainly concentrated on two types of grammatical information, case frames of verbs and prepositions attached to nouns and adjectives. Information about verb case frames and obligatory and optional prepositions is especially helpful for parsers. Using this information a parser can reduce the number of possible parsing and resolve ambiguities. This kind of grammatical information can also be used in automatic text generation. Currently most parsers use hand-generated lists of verb case frames. However, these lists are far from being complete. They do not contain the rare usages of words or specialized vocabulary.

We tried to determine the case frames of verbs, and the obligatory and optional prepositions that nouns and adjectives take. For this purpose we used a statistical database of about 180 Megabytes that is extracted from the Wall Street Journal Corpus which consists of about 38 million words. The sentences in the WSJ Corpus go through a number of processes (Figure 1.) Each sentence is first tagged by a tagger. (The tagger also works in a probabilistic manner and is trained by a tagged corpus.) The tagged sentences are parsed using a probabilistic context free grammar and the following statistics were gathered: For each word in the corpus

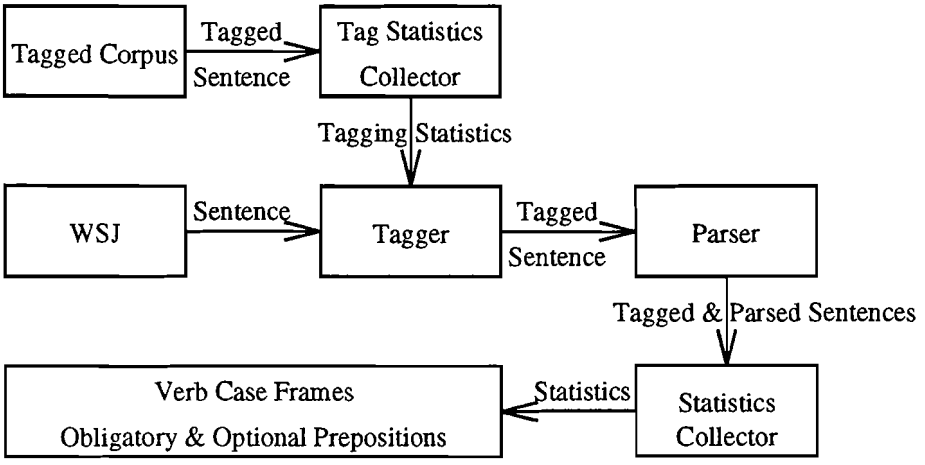

Figure 1. Statistic collection from the Wall Street Journal.

- the rules in which that word appears as the governor
- counts of these rules
- subheads of these rules

Additionally, for each preposition

- the words that the preposition is attached to

Also the list of all the rules that exist in our grammar together with the number of times they were used throughout the parsing of the WSJ corpus were generated.

In the following section the identification, filtering, and evaluation processes of verb case frames are elaborated. This is followed by a similar argument for obligatory and optional prepositions for nouns and adjectives.

2 Verb Case Frames

There have been a few attempts to find out automatically the case frames in which verbs occur in. Brent [1,2] has tried to extract this information from both untagged and tagged corpora. Similarly, Manning [4] generated case frame lists for verbs from untagged corpora.

This part of the project was mainly inspired by the study of Christopher D. Manning that was described in his paper *Automatic Acquisition of a Large Subcategorization Dictionary from Corpora.*

Verb case frames are defined as the the types of syntactic arguments a verb can take. These syntactic arguments can be in the form of objects, prepositional phrases, infinitives, etc. For example, the verb *abandon* is either followed by a noun phrase or by a noun phrase and a prepositional phrase.

Yesterday they abandoned the project. (Case frame: np)
He abandoned himself to despair. (Case frame: np_p)

Different dictionaries supply different kinds of case frames for verbs. We used the verb case frames Manning used. These were the most common case frames that appeared in almost every dictionary and linguistics literature (Table 1).

Number	Case Frame	Explanation
1	iv	Intransitive verb
2	np	Transitive verb
3	dtv	Ditransitive verb
4	that	That complement
5	np_that	Object followed by that complement
6	wh	Wh-clause complement
7	np_wh	Object followed by wh-clause
8	inf	Infinitive complement
9	np_inf	Object followed by infinitive
10	ing	-ing participle complement
11	np_ing	Object followed by -ing participle
12	adj	Adjective used as a complement
13	np_adj	Object followed by adj complement
14	adj_inf	Adj complement followed by inf
15	p	Prepositional phrase
16	np_p	Object followed by a preposition

Table 1. The sixteen verb case frames employed in our system.

2.1 Identification of Case Frames

The system first identifies the verbs, next gathers the statistics on various case frames, and finally identifies which case frames for a verb appear to be statistically significant.

The program identifies a word as a verb if it is the head of a verb rule (i.e. a rule that expands a verb phrase VB_.) This is a straightforward and simple procedure when compared with Brent's technique for verb detection where anything that occurs both with and without the suffix -ing in the text is considered as a potential verb, and if a potential verb is not preceded by a determiner or a preposition other than *to*, it is taken as a verb.

The grammar used to parse the Wall Street Journal corpus contains 1209 rules for expanding these verb phrases. These rules are mapped into our verb case frames. The verb rules that contain adverbs are treated as if there are no adverbs in the rule, e.g. the rule `VB_ -> VB ADV_ PREP` is mapped to the case frame where a verb takes a prepositional phrase, i.e., p. For the rules that contain punctuation marks or connecting words such as "and", only the part of the rule up to the punctuation mark or connecting word is taken into consideration, e.g. the rule `VB_ -> VB NN_ AND_ VB_` is classified as a transitive verb rule. Because our case frames do not have anything after a preposition, the verb rules that have nonterminals after a preposition are treated as if they have the preposition as the last nonterminal of the rule. Pronouns that appear in the rules are considered as nouns, e.g. the rule `VB_ -> VB PRON_ NN_` is mapped as the rule `VB_ -> VB NN_ NN_`, i.e., ditransitive. Additionally, some rules cannot be associated with our case frames, either because of the complexity of the rules or errors in the grammar. The probability mass of these rules, however, is less than 2% of all the rules.

For each rule used with a given verb, if we have been able to associate that rule with a particular case frame, the count for that rule is added to the count for the appropriate case frame. When all of the statistics for a particular verb have been processed, the data is filtered to determine which case frames appear to be statistically significant.

2.2 Filtering

Some of the case frames identified by our program are the actual case frames of the verb, while some are wrong ones due to a mistake in the tagger or parser or other causes, such as prepositional phrases that are not actual arguments of the verb. So the raw results have to be filtered, and the actual case frame assignments should be distinguished from the wrong ones.

The filtering method used in this program is the one proposed by Brent (1992). This method assumes that B_s is the estimated upper bound on the probability for the program to assign a wrong case frame to a verb token. Assume that a verb occurs m times in a corpus, and n of the times is classified as a certain case frame. The B_s values are used to calculate the probability that all of these n assignments are wrong. This probability is bounded by the following binomial expression:

	precision (%)	recall (%)
combined verb forms	92	52
original program	87	58
Manning's system	90	43

Table 2. Comparison with previous work.

$$\sum_{i=n}^{m} \binom{m}{i} B_s^i (1 - B_s)^{m-i}$$

In our case, n indicates the counts for each case frame, m is the total number of occurrences of the verb assigned to any case frame, and $\binom{m}{i}$ is the i-combination of m elements.

If the probability that all n assignments being false ones is low, then the probability of at least one of them being a correct one is high. So, if the above sum is less than some confidence level (in our system C = 0.02) then we assume that the case frame assignment is a correct one.

Each case frame has its own B_s value, because the probability that a case frame assignment is wrong changes from frame to frame. For example, the tagger and parser we used are more likely to make mistakes that generate extra p frames than any other frame. All the B_s values have been set empirically.

2.3 Evaluation

The Oxford Advanced Learner's Dictionary (OALD) [3] is used for testing the results of the program. The case frames that are learned by the program and the case frames in OALD do not have a direct correspondence, so OALD's 51 case frames are mapped into our 16 case frames. Then, the machine-readable version of OALD is used to extract all the verbs and their case frames automatically. That version of OALD has separate entries for the different forms of verbs (e.g., abandon, abandoned, abandoning, and abandons), so each verb form that appears in our data may be compared to the dictionary directly.

The comparison program outputs the correct case frames that our system generated, the extra ones (those that do not exist in the OALD but were generated by our program) and the missing ones (those that exist in OALD but not generated by our program) for each verb in our data.

Our system can be evaluated using two kinds of measurements. *Precision*, the ratio of the correct case frames generated by our system to the all the case frames that where generated by our system, and *recall*, the ratio of the correct case frames generated by our program to all the

case frames the dictionary supplies.

Among the most common verbs of the WSJ corpus, 30 of them were chosen randomly to be used in the evaluation. These verbs are different than those that were used by Manning, because the contexts of the corpora are different.

We achieved 87% precision and 58% recall. As mentioned previously, different forms of a verb are evaluated separately. We then combined the information about the different forms of a verb. To do this, we first grouped the different forms of every verb. Next, for each group the aggregate rule counts were calculated. This way, we tried to omit some of the incorrect case frames that were generated from only one form of the verb. This modification slightly increased the precision and resulted in a small decrease in the recall.

Table 2 compares the results of our system to those generated by Manning. It can be seen that both systems achieve almost same precision, and our system has a slightly better recall. Brent's systems, on the other hand, was able to learn only six case frames (np, dtv, that, np_that, inf, np_inf.) Table 3 shows the case frames our system generated for the group of verbs which we used in the evaluation. In this table the number of correct case frames generated by our program, the number of wrong case frames generated by our program, and the number of case frames of the verb are shown. Additionally, the final column shows what the incorrect case frames generated by our program are.

3 Prepositions

The grammatical rules of English require particular prepositions after particular nouns and adjectives. Some of the prepositions are obligatory and some are optional. Examples for these kinds of noun-preposition and adjective-preposition pairs are: "frontier between", "head of", "interested in", "glad about", etc. We tried to determine both obligatory and optional prepositions, though not the distinction between obligatory and optional.

The program that finds obligatory or optional prepositions attached to nouns starts with identifying the prepositions and nouns. We identify words that govern the noun rules as nouns, and those that govern the preposition rules as prepositions. As mentioned previously, we already have the information about which preposition is attached to which word in the corpus. Using this information, for each noun we find the prepositions attached to it and their counts. Next this raw data is filtered to get rid of the rare prepositions and errors that could have occurred during the tagging and parsing process.

Verb	# Right	# Wrong	Out of	Wrong case frames
abandon	2	0	2	
admit	3	1	7	iv
agree	2	0	5	
aim	1	0	5	
announce	2	0	3	
ask	4	0	9	
calculate	3	0	5	
decide	4	0	8	
delay	3	1	4	np_p
determine	3	0	7	
employ	1	1	2	p
engage	2	0	7	
fear	3	0	6	
gain	2	1	5	np_inf
hear	5	0	9	
join	3	0	4	
learn	3	0	7	
look	2	0	7	
make	4	0	9	
measure	3	1	4	wh
pick	2	0	3	
plunge	2	0	4	
prepare	3	0	4	
project	3	0	4	
provide	2	0	4	
retire	3	0	3	
rise	1	2	3	np, np_inf
study	3	0	6	
watch	5	0	8	
withdraw	3	0	4	
TOTAL	82	7	158	

Table 3. Comparison of verb case frame results with OALD.

	precision (%)	recall (%)
self evaluation (nouns)	87	55
dictionary evaluation (nouns)	70	45
self evaluation (adjs)	80	71
dictionary evaluation (adjs)	65	67

Table 4. Precision and recall according to the dictionary and self evaluation.

The filter that is used at this stage is the same as the one used in filtering verb case frames, i.e., the binomial filter. However, this time all prepositions are assigned the same B_s value (0.01) and the confidence level is decreased to 0.01. These values are found experimentally. For each noun, m is the number of total occurrences of that noun, and n is the number of times it appears with a specific preposition. Those that pass the filter are considered to be correct noun-preposition pairs.

The same procedure is followed to determine the prepositions for adjectives.

3.1 Evaluation

To evaluate the results of the programs, a group of nouns and adjectives are chosen randomly among the most common nouns and adjectives that take prepositions. This time however, the evaluation is done by hand because no on-line source on this type of prepositional information could be found. We used Oxford Advanced Learner's Dictionary (which does not have the information on prepositions on-line) to evaluate our results. One problem in the evaluation arose, because the dictionaries were not concise and complete in listing the prepositions that can be used with nouns and adjectives. The dictionaries give a list of prepositions that can be attached to the nouns and in the example sentences they introduce new prepositions. Also, some prepositions that sound totally correct do not appear in the dictionaries. E.g. price *of*, offer *of*, house *of*. Therefore we make two different evaluations: one only takes into account those prepositions that appears in the dictionary, and another which also considers those prepositions that sound correct. Table 4 displays the precision and recall for both nouns and adjectives. We believe that a high precision for both nouns and adjectives has been achieved. Tables 5 and 6 display our results in a small set of nouns and adjectives that were used in the evaluation. The results in these tables are generated according to the prepositions given in the dictionary and to the noun-preposition or adjective-preposition pairs that we believe are correct. In these tables the number of correct prepositions generated by our program, the number and types of wrong prepositions generated by our program, and the number of prepositions of the nouns and

Nouns	# Right	# Wrong	Out of	Wrong prepositions	Assumed correct
account	1	1	3	for	
acquisition	2	0	2		
agreement	2	0	5		on
amount	1	0	1		
bank	2	0	2		in
bid	1	0	2		
board	1	0	1		
business	2	0	4		in
chairman	1	0	1		
companies	3	1	3	that	in
control	1	0	3		
court	2	0	2		in
decline	2	0	2		
demand	2	0	3		
exchange	2	1	3	in	
group	1	0	1		
growth	2	0	2		
head	1	0	5		
index	1	0	1		
line	3	0	6		
lot	1	0	3		
market	2	0	2		
meeting	3	0	4		in, with
member	1	0	1		
number	1	0	2		
offer	2	0	2		
operations	2	0	3		in
part	1	0	2		
president	1	0	1		
price	3	0	4		in
quarter	1	1	1	from	
sale	2	0	3		in
share	2	1	3	from	
stake	1	0	2		
unit	1	1	1	in	
value	1	1	3	at	
TOTAL	47	7	86		

Table 5. Comparison of noun-preposition couples with OALD.

Adjective	# Right	# Wrong	Out of	Wrong prepositions	Assumed correct
adequate	1	0	2		
afraid	1	0	2		
available	2	0	3		
aware	1	0	2		
capable	1	0	1		
cautious	1	1	2	in	
compatible	1	0	1		
consistent	1	0	1		
different	3	0	4		in
difficult	1	0	1		
due	1	1	3	out	
eager	1	1	1	in	
enthusiastic	1	0	2		
familiar	1	0	2		
fearful	2	0	2		
guilty	1	1	2	in	
highest	3	1	3	for	of, since, in
impossible	1	0	2		
larger	2	0	2		of, than
responsible	1	0	1		
same	1	2	1	for,in	
skeptical	2	0	2		
suitable	1	0	2		
typical	1	1	1	in	
TOTAL	32	8	45		

Table 6. Comparison of adjective-preposition couples with OALD.

adjectives are shown. The prepositions that do not appear in the dictionary but were classified as correct are listed in the final column.

4 Conclusion

In this study we aimed to propose some methods to automatically extract grammatical information from untagged corpora. The project started with detecting case frames for verbs and gave promising results. Having obtained at least as good performance as the existing systems we implemented similar techniques to identify obligatory and optional prepositions for nouns and adjectives.

The accuracy of a case frame assignment for a verb or a preposition assignment for a noun or adjective heavily depends on how many times that word (noun, verb, adjective) appears in the corpus. As expected we got more precise results for more common words in the corpus.

As mentioned in the evaluation section, the dictionaries are not complete in listing all the possible prepositions that can be attached to nouns and adjectives. Our program came up with quiet many prepositions that are correct but do not appear in dictionaries, which indicates that automatic information extraction methods are a valuable complement to handed-coded lists.

References

[1] Brent, Michael R. Automatic Acquisition of Subcategorization Frames from Untagged Text. In *Proceedings of the 29th Annual Meeting of the ACL*, pages 209–214, 1991.

[2] Brent, Michael R. and Berwick, R. C. Automatic Acquisition of Subcategorization Frames from Tagged Text. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 342–345, 1991.

[3] Hornby, A. S. Oxford Advanced Learner's Dictionary of Current English. Oxford: Oxford University Press. 3rd ed., 1985.

[4] Manning, Christopher D. Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics. 1993, 235-242.