In Working Notes of the AAAI Fall Symposium on Probabilistic Approaches to Natural Language, 1992, pp. 47-53

Applying Statistical Methods to Small Corpora: Benefitting from a Limited Domain*

David Fisher†
Computer Science Division
University of California
Berkeley, California 94720
dfisher@cory.berkeley.edu

Ellen Riloff
Natural Language Processing Laboratory
Department of Computer Science
University of Massachusetts
Amherst, Massachusetts 01003
riloff@cs.umass.edu

Abstract

The application of statistical approaches to problems in natural language processing generally requires large (1,000,000+ words) corpora to produce useful results. In this paper we show that a well-known statistical technique, the t test, can be applied to smaller corpora than was previously thought possible, by relying on semantic features rather than lexical items in a corpus of limited domain. We apply the t test to the problem of resolving relative pronoun antecedents, using collocation frequency data collected from the 500,000 word MUC-4 corpus. We conduct two experiments where t is calculated with lexical items and with semantic feature representations. We show that the test cases that are relevant to the MUC-4 domain produce more significant values of t than the ones that are irrelevant. We also show that the t test correctly resolves the relative pronoun in 91.07% of the relevant test cases where the value of t is significant.

*This research was supported by the Office of Naval Research, under a University Research Initiative Grant, Contract No. N00014-86-K-0764, NSF Presidential Young Investigators Award NSFIST-8351863 (awarded to Wendy Lehnert), the Advanced Research Projects Agency of the Department of Defense monitored by the Air Force Office of Scientific Research under Contract No. F49620-88-C-0058, and ONR Contract N00014-92-J-1427.

†This research was performed while a member of the Natural Language Processing Laboratory at the University of Massachusetts at Amherst.

Introduction

The use of statistical techniques in natural language processing generally requires large corpora to produce useful results. We believe, however, that statistical techniques can be successfully applied to much smaller corpora, if the texts are drawn from a limited domain. The limited nature of the corpus may compensate for its size because the texts share common properties. Our research investigates the application of a well-known statistical technique, the t test, to the problem of resolving relative pronoun antecedents. Whereas most NLP research using statistical techniques relies on large corpora of more than one million words, we apply the t test to a much smaller corpus of only about 500,000 words. Our frequency data is based on the MUC-4 corpus, which consists of 1,700 texts in the domain of Latin American terrorism.

We apply the t test to collocations of semantic features with verbs, rather than operating solely on collocations of lexical items with verbs. The semantic features provide a generalization over the specific lexical items and, since the semantic features are specific to our domain, we can expect them to occur with greater frequencies. We find that the frequencies of the collocations of semantic features with verbs relevant to our domain are sufficient to calculate significant values of t in more than half of the test cases.

Relative Pronoun Resolution

Natural language understanding systems need to resolve the antecedents of relative pronouns. Consider sentences like the following:

> Castellar is *the second mayor* **that** has been murdered in Colombia in the last three days.

> The police department declared the state of alert as a result of a *wave of violence* **that** began with the killing of 3 people in January.

To resolve the relative pronoun **that**, a constituent from the main clause must be imported into the subclause as the subject.

We use the t test to decide which constituent from the main clause is the best candidate to be the resolvant. The data used for our frequency calculations was collected from the MUC-4 corpus [Sundheim, 1991] using the CIRCUS system [Lehnert, 1990; Lehnert *et al.*, 1991a; Lehnert *et al.*, 1991b]. Each clause was parsed into a subject-verb-object (**SVO**) triple, storing the morphological root and semantic features for each of the subject (**S**), verb (**V**), direct object (**O**) in the **SVO** database. Passive voice constructions were transformed into an active representation for the frequency calculations.

An alternative approach to the resolution problem for *who*, involving learning resolution heuristics using the MUC-4 corpus, is presented in [Cardie, 1992a; Cardie, 1992b]. [Dagan and Itai, 1991] used collocational frequency data from a 150 million word corpus to resolve the pronoun *it*, however only 36 of their 74 test cases requiring resolution of *it* exceeded the minimum threshold needed to apply their statistical filter.[1]

Before presenting the results of our experiments, we first describe the MUC-4 corpus and the design of the experiments. This includes a description of the t test, the frequency data used for its computation, and the test sets from which we drew our test cases. We analyze the distinction between the relevant and irrelevant test cases and significant t values using $\chi 2$.

The MUC-4 Corpus

The Fourth Annual Message Understanding Conference (MUC-4) was sponsored by DARPA in 1992 to provide a basis for evaluating the state of the art in English language text analysis systems. To perform these evaluations, 1,700 texts were selected by keyword search from a database[2] containing newswire stories from around the world. The search extracted texts that were concerned with terrorist activities, such as bombings, kidnappings, and murders. The texts are unconstrained, containing original language, misspellings, ungrammatical statements and the like.

The MUC-4 corpus contains approximately 500,000 words, with 18,500 unique lexical items. The domain is terrorism in Latin America; a text is considered to be *relevant* to the domain if it concerns one or more terrorist events occurring in a specific set of Latin American countries. However roughly half of the texts in the corpus are *irrelevant* because, although they contain a keyword associated with terrorism, they do not reference any specific terrorist events.

We differentiate individual sentences in a similar fashion. A relevant sentence generally contains a verb, such as kill, kidnap, or explode, that describes a relevant event. Sentences that do not reference a terrorist event can be classified as irrelevant. This relevance distinction is fundamental to the design of our experiments.

Experimental Design

We conducted two experiments to evaluate the performance of the t test on relative pronoun resolution using a set of 196 sentences extracted from 200 texts from the MUC-4 corpus. These 200 texts were not among those used to collect the collocation frequency data. The test sentences contained every occurrence of **that**, **which**, and **who** that were used as relative pronouns in these 200 texts. Seven of the clauses extracted from the test sets were discarded prior to the experiments because of parser errors that caused either the antecedent or the relative pronoun to be parsed as the wrong part of speech. The remaining 219 clauses contained 70 instances of **that**, 65 of **which**, and 84 of **who**.

Our objective was to determine if the limited domain would provide sufficient collocation frequencies to calculate significant values of t. Our experiments use the frequency data from 1,500 of the MUC-4 texts to resolve a relative pronoun by choosing the antecedent that shows the strongest association with the verb in the embedded clause. For each test case we compute t for each of the candidate antecedents, paired with the main verb from the embedded clause. We make the assumption that the candidate belongs in the subject role of the embedded clause, e.g. for active clauses the collocation would be **SV**, and for passive clauses **VO**. This simplifying assumption held for all of the test cases; there were no instances where the antecedent did not fill the role of subject in the embedded clause.

We handle passive constructions by transforming them into an active representation. The subject of a passive clause is assigned to the object (**O**) category for the purposes of frequency calculations. For example, in the sentences

The FMLN killed *three peasants* today.
Three peasants were killed by the FMLN today.

the *three peasants* are the ones being killed. So when we try to resolve a passive clause like

[1] Although the selection of the minimum value was arbitrary, it provides a reasonable approximation of the notion of *significance* as we are using it here.

[2] This database has been constructed and is maintained by the Foreign Broadcast Information Service (FBIS) of the U.S. Government [Sundheim, 1991].

We remember the murder of the *Jesuit priests* **who** were killed ...

we compute t to select the candidate that is most likely to be the object of *killed* in an active construction.

The first experiment uses the morphological roots of the lexical items for the computation of the frequencies. The decision to use the morphological roots was based on the concern that the frequencies of lexical items would be far too low to calculate meaningful values of t. It was expected that the relatively low frequencies of some of the words would make it difficult to produce significant results.

The second experiment uses the semantic features of the nouns instead of the specific lexical items; this evaluates performance on the equivalence classes created by the semantic features. The semantic feature hierarchy used by CIRCUS in MUC-4 contains 67 features and was based on the hierarchy suggested by the MUC-4 task description. Lexical items can have zero or more semantic features. By collapsing the actual words into equivalence classes, we try to overcome the problem of frequencies that are too small to calculate t. In addition, we believe that t values based on semantic features are more appropriate for selecting the antecedent because they show preference for the most likely meaning, as opposed to the most likely word.

t Test

The t test is based on the probabilities of items occurring in collocation. t provides a measure of the likelihood that two items will occur together. [Church *et al.*, 1991] has shown the t test to be an effective tool when working with frequency data from a corpus of over one million words.

To use the t test as a control strategy for resolving relative pronouns, we first compute t for each of the candidate antecedents from the main clause paired with the verb of the subclause and then consider only those candidates which produced a *significant* value of t. In the cases where more than one candidate generated a significant t value, the one with the largest t value was selected. In the cases where two or more candidates had equal, significant, t values, the rightmost candidate (the one closest to the relative pronoun) was chosen as the antecedent.

For these experiments we use two significance levels for t, $t > 1.65$ which is significant at the $p < 0.05$ level and $t > 1.27$ which is significant at the $p < 0.10$ level. The expression $p < 0.x$ indicates that there is a less than $x\%$ probability that the collocation being tested occurred due to chance. The results we are presenting here were produced at the $p < 0.10$ level. The use of the $p < 0.10$ significance level added many additional cases (5 for the lexical items and 16 for the semantic features). In every case the correct antecedent was identified. The summary results tables for the $p < 0.05$ level are given in the Appendix.

SVO Database

Our experiments use **SVO** frequency data from the MUC-4 corpus for computing t. We constructed the **SVO** database[3] by parsing 1,500 texts from the MUC-4 corpus using CIRCUS. CIRCUS is a conceptual sentence analyzer that produces a semantic case frame representation of a sentence. It uses a stack-oriented control structure for syntactic processing, a marker-passing mechanism for predictive preference semantics, and bottom-up prepositional phrase attachment. It does not construct a full syntactic parse tree of the input. CIRCUS parses every input clause, but only produces case frame instantiations for some of them.

CIRCUS can be characterized as a text-skimmer that uses *selective concept extraction* to extract relevant information from a sentence. It relies on a domain-specific dictionary of *concept nodes*, which are case frames activated by relevant words or phrases in the text. Once activated, concept nodes pick up relevant information from the sentence and store it as slot fillers inside case frames. For relevant sentences, CIRCUS generates one or more concept nodes that contain the relevant information. Since concept nodes are defined only for relevant events and items, CIRCUS generates no concept nodes for irrelevant sentences.

For the purposes of this research we used only the syntactic processing and semantic feature components of CIRCUS. The CIRCUS lexicon used for MUC-4 contains roughly 6,500 lexical entries. The lexicon was augmented for this research with an additional 4,200 entries generated from Moby™ Part-of-Speech, which provided part of speech tagging, and with 900 proper nouns that were marked with the semantic feature `proper-name`. We used demand-driven file lookup to generate the entries from Moby™ Part-of-Speech. This was done by parsing the MUC-4 corpus with CIRCUS and looking up each undefined word, and its morphological roots, in Moby™ Part-of-Speech. The words that were not found in Moby™ Part-of-Speech were assumed to be proper nouns, and were assigned the semantic feature `proper-name`.

These 1,500 texts produced 56,475 **SVO**s. Morphological roots were used for the verbs and nouns when such roots where available via CIRCUS' morphological

[3]The **SVO** database was constructed using RTM, a common lisp Relational Table Manager, which was developed by Paul Silvey of the Experimental Knowledge Systems Laboratory at the University of Massachusetts at Amherst.

Verbs	56,475
Lexical Items	
Verb only	1,831
Subjects	42,505
Objects	33,939
Subject and Object	21,800
Semantic Features	
Verb only	3,098
Subjects	40,561
Objects	32,770
Subject and Object	19,954

Table 1: **SVO** Database

analysis. The frequency breakdown of the constituents is shown in Table 1 (a clause had to have a verb in order for it to be stored in the database). In the table, *Verb only* is the number of clauses that had no subject or object. *Subjects* is the number of clauses containing a subject with or without an object, and *Objects* is the number of clauses containing an object, with or without a subject. *Subject and Object* is the number of clauses which had both a subject and an object.

Passive voice clauses were transformed into an active voice representation prior to being inserted into the database. This was done by assigning the subject of the clause to the object category (**O**), and by assigning the object of the preposition *by*, when it was present, to the subject (**S**) category.

Test Sets

The test clauses were drawn from the remaining 200 texts in the MUC-4 corpus. These were the texts used for the MUC-4 final evaluation, labelled TST3 and TST4. Although they share the terrorism domain, there is a major difference between the two sets. TST3 was drawn from the same time period as the 1,500 texts that were used to construct the **SVO** database.

TST4, however, was drawn from a time period approximately three years earlier. In this respect TST4 is not as well represented in the database as TST3 at the lexical level. That is to say, different people are being murdered, different physical targets are being attacked, and different terrorist organizations are involved. At the semantic feature level, on the other hand, the same classes are involved in the events in both test sets.

Experimental Results

Prior to running our two experiments, we tested the collocations of verbs relevant to the MUC-4 domain with nouns to see if we would obtain significant values of t. The results of these tests proved to be interesting. We expected to see strong associations between verbs and nouns that often occur together in terrorist event descriptions. Table 2 shows some examples of these types of associations for the verbs *found*, *murder*, and *kill*.

Things that are *found*

w	$t(\text{found}, w)$
body	4.5440
weapon	2.4892
corpses	1.7215

Things that are *murdered*

w	$t(\text{murder}, w)$
people	2.2288
priest	1.6927
civilian	1.6826

Things that are *killed*

w	$t(\text{kill}, w)$
people	5.3317
soldier	5.2629
guerrilla	4.0448
priest	3.9310

Things that *kill*

w	$t(w, \text{kill})$
guerrilla	2.5449
soldier	2.5059
terrorist	2.1044

Table 2: Associations in the MUC-4 Corpus

Tables 3 and 4 illustrate the power of using semantic feature representations for calculating t. They show t values for some of the semantic features that inherit from `weapon` (e.g., weapon) and `human` (e.g., body & corpses). We see that many different kinds of weapons are likely to be *found* at the semantic feature level. Yet not all of the lexical items that have a significant t value for their semantic features have a significant t value for the word itself. For example *gun* has a semantic feature which inherits from `weapon`, `gun`, but $t(\text{found}, gun) = 0.9817$ is an insignificant value of t.

It is also not the case that when t is significant for a lexical item with a verb that it is necessarily also significant for that item's semantic features with that same verb. Consider the following two nouns with the verb *state*.

w	$t(\text{found}, w)$
`rifle`	3.8269
`machine-gun`	3.7053
`handgun`	3.5836
`gun`	3.4470
`bomb`	3.2565

Table 3: Isa `weapon`

w	$t(\text{found}, w)$
`clergy`	3.8148
`civilian`	3.7902
`legal-or-judicial`	3.5513
`govt-official`	3.6449
`active-military`	3.2362

Table 4: Isa `human`

	Relevant		Irrelevant		Total	
Cor.	22	23.91%	14	11.02%	36	16.44%
Inc.	0	0.00%	1	0.79%	1	0.46%
Ins.	70	76.09%	112	88.19%	182	83.10%
Tot.	92		127		219	

Table 5: Lexical Items $p < 0.10$

w	$t(w, \text{state})$
report	4.9231
{`communications`, `media`}	0.5155
bulletin	1.6779
{`entity`, `organization`}	2.4190

report and **bulletin** both generate significant t values in conjunction with *state*, where **report** is preferred. The semantic features produce the opposite result; the features for **report** are not significant so the features for **bulletin** are preferred. These two nouns do not appear as competing candidates in the test cases, but if they did, both the lexical item and semantic feature tests would generate a significant response, and the two responses would disagree. In three of the test cases the lexical items produced significant t values although the semantic features did not.

Lexical Items

Our first experiment uses the morphological roots of the lexical items to compute collocation frequencies. Table 5 shows the performance of the t test on the resolution task, broken down by relevance to the MUC-4 domain.[4] When we consider only the cases where the value of t is significant, we find that 100.00% (22/22) of the relevant and 93.33% (14/15) of the irrelevant cases are correctly resolved with the t test. We also find that there are twice as many instances, on a percentage basis, of significant t values for the relevant cases as for the irrelevant cases.

A number of interesting observations emerged from the lexical item experiment. First, there appears to be a relationship between relevance and significant t values, given the percentage difference in the frequencies. To test the relationship we apply the χ^2 test to this 2×2 contingency table.

	Relevant	Irrelevant
Significant t	22	15
Insignificant t	70	112

This produces $\chi^2 = 4.7366$ which is significant at the $p < 0.0295$ level, giving evidence that there is a meaningful relationship between relevance and significance.

In addition, we found considerable distinctions between the two test sets (see Table 6). For TST3 (the set from the same time period as the sets used to construct the **SVO** database) we find almost three times as many significant instances for the relevant clauses as the irrelevant clauses. In addition, this set also contains the single incorrect case. Twice as many significant t values were produced for TST3 as for TST4 for the combined relevant and irrelevant cases.

This is consistent with the differences between the two sets. Recall that TST4 comes from an earlier time period than TST3, so many of the lexical items in the relevant events are different. Appealing to χ^2 again with the following table we have $\chi^2 = 3.6374$ which is significant at $p < 0.0560$.

	TST3					
	Relevant		Irrelevant		Total	
Cor.	18	33.96%	11	11.96%	29	20.00%
Inc.	0	0.00%	1	1.09%	1	0.69%
Ins.	35	66.04%	80	86.96%	115	79.31%
Tot.	53		92		145	
	TST4					
	Relevant		Irrelevant		Total	
Cor.	4	10.26%	3	8.57%	7	9.46%
Inc.	0	0.00%	0	0.00%	0	0.00%
Ins.	35	89.74%	32	91.43%	67	90.54%
Tot.	39		35		74	

Table 6: Lexical Items by Test Set $p < 0.10$

[4] In each of the tables the percentages are based on the total number of test cases. The following abbreviations are used throughout: **Cor.** Number of correctly resolved antecedents. **Inc.** Number of incorrectly resolved antecedents. **Ins.** Number of cases with insignificant values of t. **Tot.** Total for each column.

	TST3	TST4
Significant t	30	7
Insignificant t	115	67

This significance level indicates that the relationship between significance and test set is not quite meaningful[5], however we feel that it is close enough to justify the distinction at the lexical level.

It is also interesting to see where the t test failed to select the correct antecedent. The single resolution error in TST3 occurred in the processing of **who** in the following sentence, where the correct antecedent is in *italics* and the incorrect selection is underlined.

> Salvadoran president Alfredo Cristiani stated today that it would be no problem for *the fourth penal judge*, **who** is in charge of the case of the Jesuits murdered by a group of soldiers, to order that cadets who are taking courses abroad return to the country.

Applying the t test to the verb *is* (TO-BE) produces

$$t(\text{problem}, \text{TO-BE}) = 3.3294$$
$$t(\textit{judge}, \text{TO-BE}) = -2.2868$$

selecting problem as the antecedent. TO-BE is difficult for frequency based statistical techniques, such as the t test, because it both occurs frequently and does not provide discrimination of its subjects or objects. In a sense, *anything* can be the subject, or object, of a clause where the main verb is TO-BE, and that is about all we can know about it. [Dagan and Itai, 1991] excluded all instances where the main verb was TO-BE from their experiments.

Overall we find that we can rely on the t test in 23.91% of the relevant test cases, and 16.90% overall, when using the frequency data for the lexical items. This is not often enough for us to feel confident about using the t test as a general method for resolving relative pronouns. However the difference between the relevant and irrelevant cases does illustrate the effect of using a corpus with a limited domain. To reap the maximum benefit from this effect, we turn to the semantic feature representations.

Semantic Features

The results for the semantic features experiment are shown in Table 7, with the breakdown by test set in Table 8. As with the lexical items, we find that more of the relevant (91.07%, 51/56) as opposed to the irrelevant (84.21%, 48/57) cases are correctly resolved when the t values are significant.

[5]We accept the hypothesis that there is no relationship between the two for $\alpha = 0.05$. We would need a value of $\chi^2 > 3.84$ in this case.

	Relevant		Irrelevant		Total	
Cor.	51	55.44%	48	37.80%	99	45.21%
Inc.	5	5.43%	9	7.09%	14	6.39%
Ins.	36	39.13%	70	55.11%	106	48.40%
Tot.	92		127		219	

Table 7: Semantic Features $p < 0.10$

	TST3					
	Relevant		Irrelevant		Total	
Cor.	33	62.26%	37	40.22%	70	48.28%
Inc.	3	5.66%	4	4.35%	7	4.83%
Ins.	17	32.08%	51	55.43%	68	46.90%
Tot.	53		92		145	
	TST4					
	Relevant		Irrelevant		Total	
Cor.	18	46.15%	11	31.43%	29	39.19%
Inc.	2	5.13%	5	14.29%	7	9.46%
Ins.	19	48.72%	19	54.29%	38	51.25%
Tot.	39		35		74	

Table 8: Semantic Features by Test Set $p < 0.10$

In this case there are a third again as many relevant instances as irrelevant. Here $\chi^2 = 4.8390$, which is significant at $p < 0.0278$, using the following table.

	Relevant	Irrelevant
Significant t	56	57
Insignificant t	36	70

So for both the lexical items and the semantic features we see that the performance of applying the t test to the resolution task, and the ability to generate significant t values, is better for the test cases that are relevant to the MUC-4 domain.

The distinction between the two test sets does not hold for the semantic features. $\chi^2 = 0.2314$, which is significant at $p < 0.6305$.

	TST3	TST4
Significant t	77	36
Insignificant t	68	38

At the semantic level we can not differentiate the two test sets in terms of the number of significant t values they produce. This is as expected since we use the same semantic feature hierarchy for both test sets. The same classes appear as the complements of the relevant verbs in both time periods, only some of the words are different.

Using semantic features produced more significant results than for lexical items, and it also produced more errors. There were 14 cases where the antecedent chosen was incorrect, 10 involving **who** and four involving

that. Six of the 14 errors, three from **who** and three from **that**, are in clauses where the main verb is TO-BE. The semantic feature `entity`, the root node of the hierarchy, produces the most significant value of t for this verb.

Seven of the 10 **who** errors could have been prevented by adding the constraint that the resolvant of **who** must be a member of, or inherit from one of the members of, the set {`human`, `organization`, `proper-name`}. The remaining three **who** errors are cases where both the correct antecedent and the incorrect selection satisfied this constraint, and the more general candidate was preferred.

An example of this type of error appears in the following sentence:

> Another report indicates that the authorities have identified the corpse of *policeman David Diaz Cortez*, **who** was killed at daybreak today in a skirmish by rebels against Suchitoto military.

	w	$t(\text{kill}, w)$
corpse:	{`human`}	7.8922
Cortez:	{`proper-name`, `law-enforcement`}	4.5616

The number of occurrences of `proper-name` with other verbs lowers the value of t. This is partially the result of a CIRCUS heuristic which assigns the semantic feature `proper-name` to undefined words, and it is also because many non-`human` things, such as locations and organizations, have the feature `proper-name`.

Applying the t test to semantic features produces significant results in 60.87% of the relevant cases, and 51.60% overall. In those cases the t test correctly identifies the antecedent for 91.07% of the relevant clauses, and 87.61% overall. If we add the candidate constraint for resolving **who** with semantic features, the overall performance increases to 93.81% (106/113), with 94.74% (54/57) for the relevant cases and 91.23% (52/57) for the irrelevant ones.

Conclusion

We set out to examine the possibility of applying a statistical technique, the t test, to the problem of resolving the antecedents of relative pronouns, using frequency data from a small corpus drawn from a limited domain. Although we have evaluated its performance on the resolution task, our primary concern is the ability to generate significant values of t using the **SVO** database.

We find that, in each experiment, more significant results are produced for the test clauses that are relevant to the MUC-4 domain. The χ^2 tests of the relationship between relevance and significant t values indicates that this relationship is significant for both the lexical items and the semantic features. This is consistent with our expectation that the limited nature of the domain would compensate for the size of the corpus. This provides evidence that these types of statistical techniques may perform well when applied to other similarly limited domains.

The performance of the t test using semantic features on the relevant test sentences shows that this type of statistical technique can be applied to smaller corpora than was previously thought possible by taking advantage of the limited nature of the domain. In 60.87% of the test cases we correctly identify the antecedent 91.07% of the time. For the lexical items we correctly identify every antecedent in 23.91% of the test cases.

From this we conclude that, at the lexical item level, frequency data from a larger corpus is necessary for this application of the t test. The number of cases with significant t values is too low to justify using the t test as a control strategy for pronoun resolution. Although there is an effect from using a limited domain, it is not enough.

At the semantic feature level the number of significant cases seems to be sufficient to justify using the t test, leaving the remaining cases to some other heuristic. Performance on the pronoun resolution task, however, is not a large improvement over a simple heuristic that selects the preceding phrase as the antecedent. [Cardie, 1992b] measured performance of 86% correct resolutions for the preceding phrase heuristic when resolving **who**. So it appears that the task of resolving relative pronouns may not be the most compelling application of statistically-based sentence analysis techniques.

Large corpora are necessary to construct the frequency data for statistical techniques, such as the t test, when they are applied to texts from an unconstrained domain. Smaller corpora, however, may be sufficient to generate useful frequency data if they are drawn from a limited domain. These results show that this holds for the MUC-4 corpus, which is less than half the size of the corpora that have been used by others in previous research. In each case the results indicate that further exploration, with other frequency based statistics and other problems, is needed to determine how much utility can be derived from using a small corpus with a limited domain.

Appendix

	Relevant		Irrelevant		Total	
Cor.	18	19.57%	13	10.24%	31	14.16%
Inc.	0	0.00%	1	0.79%	1	0.46%
Ins.	74	80.43%	113	88.98%	187	85.39%
Tot.	92		127		219	

Lexical Items $p < 0.05$

	Relevant		Irrelevant		Total	
Cor.	41	44.57%	42	33.07%	83	37.90%
Inc.	5	5.43%	9	7.09%	14	6.39%
Ins.	46	50.00%	76	59.84%	122	55.71%
Tot.	92		127		219	

Semantic Features $p < 0.05$

References

Cardie, C. 1992a. Corpus-Based Acquisition of Relative Pronoun Disambiguation Heuristics. In *Proceedings, 30th Annual Meeting of the Association for Computational Linguistics*, Morristown, NJ. Association for Computational Linguistics. 216–223.

Cardie, C. 1992b. Learning to Disambiguate Relative Pronouns. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, Menlo Park, CA. AAAI Press.

Church, K.; Gale, W.; Hanks, P.; and Hindle, D. 1991. Using Statistics in Lexical Analysis. In Zernick, U. (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Dagan, I. and Itai, A. 1991. A Statistical Filter for Resolving Pronoun References. In Feldman, Y. A. and Bruckstein, A. (eds.), *Artificial Intelligence and Computer Vision*. Elsevier Science Publishers, North-Holland. 125–135.

Lehnert, W.; Cardie, C.; Fisher, D.; Riloff, E.; and Williams, R. 1991a. University of Massachusetts: Description of the CIRCUS System as Used for MUC3. In *Proceedings of the Third Message Understanding Conference*, San Mateo, CA. Morgan Kaufmann Publishers, Inc.

Lehnert, W.; Williams, R.; Cardie, C.; Fisher, D.; and Riloff, E. 1991b. The CIRCUS System as Used in MUC3. Technical Report 91–59, University of Massachusetts at Amherst, Amherst, MA.

Lehnert, W. 1990. Symbolic/Subsymbolic Sentence Analysis: Exploiting the Best of Two Worlds. In Barnden, J. and Pollack, J. (eds.), *Advances in Connectionist and Neural Computation Theory*. Ablex Publishers, Norwood, NJ.

Sundheim, B. 1991. Overview of the Third Message Understanding Evaluation and Conference. In *Proceedings of the Third Message Understanding Conference*, San Mateo, CA. Morgan Kaufmann Publishers, Inc.