Detecting Word Substitutions in Text

SW. Fong, D. Roussinov, D.B. Skillicorn *Member, IEEE*

***Abstract*— Searching for words on a watchlist is one way in which large-scale surveillance of communication can be done, for example in intelligence and counterterrorism settings. One obvious defense is to replace words that might attract attention to a message with other, more innocuous, words. For example, the sentence "the attack will be tomorrow" might be altered to "the complex will be tomorrow", since 'complex' is a word whose frequency is close to that of 'attack'.**

Such substitutions are readily detectable by humans since they do not make sense. We address the problem of detecting such substitutions automatically, by looking for discrepancies between words and their contexts, and using only syntactic information. We define a set of measures, each of which is quite weak, but which together produce per-sentence detection rates around 90% with false positive rates around 10%. Rules for combining per-sentence detection into per-message detection can reduce the false positive and false negative rates for messages to practical levels. We test the approach using sentences from the Enron email and Brown corpora, representing informal and formal text respectively.

***Index Terms*— textual analysis, counterterrorism, word frequencies, data mining, pointwise mutual information, co-occurrence.**

I. Motivation

Groups that are involved in illicit acts, whether terrorists or criminals, must communicate with one another. They are surely aware of the possibility that both the existence of their communications, which provides evidence of their links to one another, and the content of their communications, which provides evidence of their thinking and actions, are targets for intelligence or law enforcement.

How they attempt to conceal their communication depends on what kind of interception is being done. If interception is being done for particular senders and receivers, then only the content of the messages can be concealed, and encryption may be the technique of choice. If interception is being done by automated scanning of large numbers of messages, for example by government intelligence programs or organizational analysis of email, obfuscation of content may be a better technique. In this setting, encryption draws attention to a message that might otherwise not be noticed.

The first level of analysis in widespread message interception is to scan for the presence of words from a list of significant words, a watchlist. Messages that contain such words may be selected for further analysis, for example using more-powerful text-mining algorithms, or even human analysis. Obfuscation replaces words that are, or may be, on the watchlist by more innocuous words, making the message seem more 'ordinary', and so likely to be unselected for further analysis.

The form of the substitutions depends on whether the screening process is done by humans or by software. For example, al Qaeda was, for a time, using the word 'wedding' in place of the word 'attack'. This substitution is obviously aimed at human readers, since many of the things that might be said about an attack might also be said about a wedding: they both happen at particular places and times, and require coordination to make sure that all the participants arrive at the proper time.

Humans use semantic and deep contextual information to judge whether a substitution has occurred. Software is limited to surface properties, for example frequencies of words and strings of words. Avoiding detection by software may require paying more attention to such properties. For example, 'attack' is the 1072nd most common word in English (according to a list at `www.wordcount.org/main.php`), while 'wedding' is the 2912nd most common word, creating the possibility of detecting the example al Qaeda substitution automatically on the basis of frequency differences.

When such substitutions must be made 'on the fly', for example during phone calls and perhaps emails, and particularly under conditions of stress, it is plausible that the choices for replacement words may be poor, and so the presence of a substitution easily detectable [1]. However, word

SW. Fong and D.B. Skillicorn are with Queen's University, Kingston, Canada K7L 3N6

D. Roussinov is with Arizona State University, Tempe, USA

frequency information is readily available, so it is possible that, in ordinary circumstances, a terrorist or criminal group might adopt a standard set of substitutions, in which words they do not wish to use are replaced by other words *with similar frequencies*. This paper addresses the problem of detecting such substitutions.

As recent experience with popular web sites such as MySpace.com indicates, the electronic communication facilitated by those sites may provide leads to detecting and thwarting violent plots [2]. Since much of that communication is publicly available, perpetrators may choose to replace words that might attract attention. It is also known that Salafist terrorist use web sites for internal communication, and for information dissemination. Posting online content that can facilitate terrorist acts, such as posting instructions on how to make explosives, poisons, and so on has recently been made illegal and techniques to detect such malevolent content have been explored [3]. The authors of such materials may try to avoid certain eye-catching words, replacing them with more ordinary ones.

The contribution of this paper is the design of a set of measures that can be applied to sentences, and whose values are predictive of the presence of a substituted word. Each of these measures is relatively weak on its own. However, each makes errors on different kinds of sentences, so combining them into ensembles produces substitution detectors with high accuracies. Detection rates of around 90% with false positive rates around 10% are achieved. We demonstrate using sentences drawn from the Enron email corpus and the Brown news corpus.

II. Related Work

A standard model for many natural language problems is to assume a language-generation model that describes how sentences in English are generated, and an alteration model that describes how such sentences are changed in the problem domain being considered. The probability of a given sentence w being generated is given by some probability $P(w)$. The alternation model changes w to some new sentence y with probability $P(y \mid w)$. The task is to estimate w given y [4].

In the problem we address, the alteration model is the replacement of some set of words with other words of similar frequency. We are interested, not in predicting the original sentence (which would be extremely difficult), but in detecting when $P(w)$ differs significantly from $P(y)$. Some early results have already appeared [5].

An easier variant, the problem of detecting a substituted word with substantially different frequency from the word it replaces was addressed by Skillicorn [1]. This work considered, not individual sentences, but large collections of messages. The existence of identical substitutions in different messages was shown to be detectable, via the correlations that were created among them, using matrix decompositions.

Speech recognition uses an alteration model in which text is converted to an analogue wave form. Predicting the original sentence w is done using the left context of the current word and a statistical model of word co-occurrences. Such algorithms are heavily dependent on left-to-right processing, backing up to a different interpretation when the next word becomes sufficiently unlikely [6]. Speech recognition differs from the problem addressed here because it is limited to the left context, whereas we are able to access both left and right contextual information. Further, speech recognition techniques must be lightweight because of the need for near realtime performance.

Detecting misspellings uses an alteration model that incorporates common keystroke errors, themselves derived from visual, aural, and grammatical error patterns [7]. This problem differs from the problem addressed here because misspelled words are easily distinguishable from ordinary words, and because the alterations are of limited forms.

Spam detection is closer to our problem in the sense that the alteration model assumes human-directed transformations with the intent to evade detection by software. For example, SpamAssassin uses rules that will detect words such as ‘V!agra’. The problem is similar to detecting misspellings, except that the transformations have properties that preserve certain visual qualities rather than reflecting lexical formation errors. Lee and Ng [8] detect word-level manipulations typical of spam, using Hidden Markov Models. They addressed the question of whether an email contains examples of obfuscation by word substitution, expecting this to be simpler than recovering the text that had been replaced. They remark that detecting substitution at all is ‘surprisingly difficult’ [8, Section 5] and achieve prediction accuracies of around 70% using word-level features.

The task of detecting replacements can be considered as the task of detecting words that are "out of context," which means surrounded by the words with which they typically do not co-occur. The task of detecting typical co-occurrences of words in specific contexts was considered in [9, 10].

Using Google (or other Internet search engines with large coverage) to check for spelling and grammatical errors has been suggested in the academic literature [11]. Indeed, since substitutions frequently result in incorrect grammatically or semantically formed phrases, detecting such errors may also detect substitutions. For example, the erroneous use of a word in the phrase "had ice-cream for desert" means that it occurs on the Web only 44 times, according to Google. The correct phrase "had ice-cream for dessert" occurs 316 times. However, no evaluation was performed in [11] and we are not aware of any other formal studies in this direction.

III. Measures

We expect that a substituted word creates an anomaly in the flow of a sentence because its meaning does not fit with the meanings of the words around it in the sentence. It was selected on the basis of a much shallower property, its overall frequency in English.

If the substituted word's frequency is almost the same as that of the word it replaces, then we must look at elements of the context in which it appears to find ways to detect its presence. For a human this is often easy, since a word of equivalent frequency is unlikely to make sense in context. However, sometimes, especially for common words with multiple meanings, replacement can be difficult even for humans to detect. For example, the words 'results' and 'conditions' have almost the same frequency, so altering the sentence "the results are quite poor" to "the conditions are quite poor" is difficult to detect either semantically or syntactically.

We wish to detect substitutions of equally frequent words without direct semantic information. An obvious starting point is to consider the frequencies of pairs of words (2-grams). If we consider all of the 2-grams of a sentence, then a substituted word appears as a member of two adjacent 2-grams. We might expect that the frequencies of these 2-grams would be lower than those of the 2-grams around them because these particular pairs of words do not belong together semantically; and this is reflected in their observed frequencies. 2-grams consider small contexts on either side of the substituted word.

Unfortunately, Ferrer i Cancho and Solé [12] have shown that the graph of English word adjacencies has a small-world property. Words can be imagined as occurring in a layered sphere, with very common words near the center and rare words towards the outside. Their result implies that any word is almost always both preceded and followed by a very common word. In our setting, this means that the immediate context of a substituted word is likely to tell us little about how well that word 'fits' into a sentence. Measures that are able to consider much larger contexts are needed. Of course, we could consider 3-grams, 4-grams, and so on to get information about larger contexts.

We use large text repositories as oracles for the natural frequency of words, bags of words, and strings. It has been observed that even relatively short strings do not occur verbatim, even in the largest text repository. For example, Zhu and Rosenfeld [13] noted that about ten percent of 3-grams from fresh news stories did not already appear in several search engines. Hence it is likely that we can get no frequency information about many strings of words of length 3 or longer.

We will use the following sentence as a running example: "we expect that the attack will happen tonight". As expected, this exact sentence occurs with zero frequency at both Yahoo and Google, even though it is not a particularly surprising sentence. In this sentence, the word 'attack' is the word mostly likely to single this sentence out for further analysis. A word with similar frequency to 'attack' in English is 'campaign', so the sentence with a substitution we will consider is "we expect that the campaign will happen tonight". (Note that the sentence with the substitution is not semantically more surprising than the original sentence in this particular case.)

We have designed and adapted a set of measures that view the relationship between a word and its contexts in different ways. Some of these are based on the frequencies of short strings, while others treat sentences and strings of words as bags of words. Here are the measures:

a) Sentence Oddity (SO): This measure considers a sentence as a whole, and the relationship between the entire sentence, and the sentence with a particular word of interest deleted. As noted above, we can only get useful frequency estimates by treating sentences as bags of words, that is we generate search engine queries with a list of the words in the sentence, rather than treating the entire sentence as a quoted string.

Sentence oddity is based on the observation that removing a contextually appropriate word from a sentence should not substantially change the frequency of the resulting bag of words in comparison to the frequency of the entire sentence, since the contextually appropriate word co-occurs frequently with the other words in the sentence. On the other hand, removing a contextually inappropriate word might be expected to produce a large increase in frequency of the remaining bag of words because it would only rarely co-occur with the other words. Hence we define the sentence oddity of a sentence with respect to a particular target word as:

$$\text{SO} = \frac{\text{frequency of bag of words, target word removed}}{\text{frequency of entire bag of words}}$$

Sentence oddity should be large for a sentence in which a word has been substituted.

The frequency, at Yahoo, of our example sentence with ‘attack’ removed is 5.78M, while the frequency of the entire sentence is 2.42M, so the sentence oddity of the example sentence is 2.4 (=5.78/2.42). For the sentence with the substitution, the frequency of the entire sentence is 1.63M so the sentence oddity is 3.5 (=5.78/1.63). As expected, the sentence oddity of the sentence containing the substitution is significantly larger than that of the original sentence.

b) Enhanced Sentence Oddity (ESO): The numerator in the sentence oddity measure includes some sentences that contain the word being considered; that is the numerator counts some sentences that are also counted in the denominator. It is useful to define enhanced sentence oddity in which the numerator explicitly excludes the word being considered. Hence we define the enhanced sentence oddity of a sentence with respect to a particular target word as:

$$\text{ESO} = \frac{\text{frequency of bag of words, target word excluded}}{\text{frequency of entire bag of words}}$$

Again, enhanced sentence oddity should be large for a sentence in which a word has been substituted.

For the example sentence, the frequency of the numerator is 3.36M, giving an enhanced sentence oddity of 1.4 (=3.36/2.42). For the sentence with the substitution, the frequency of the numerator is 4.14M, giving an enhanced sentence oddity of 2.5 (=4.14/1.63).

TABLE I
EXAMPLE K-GRAMS

“the attack will happen”	$f = 489$
“the attack will happen tonight”	$f = 1$
“that the attack will happen”	$f = 204$
“expect that the attack will happen”	$f = 0$
“the campaign will happen”	$f = 26$
“the campaign will happen tonight”	$f = 0$
“that the campaign will happen”	$f = 0$

c) k-gram frequencies (k-GRAM): The difficulties of using the frequencies of exact strings containing the word of interest are illustrated by looking at the frequencies of substrings of our example sentences. These are illustrated in Table I.

Frequencies overall are lower for the fragments of the sentence that contains the substitution, but we would not, in practice, know the frequencies of the original sentence to compare them with. Frequencies of exact strings are often so low that they are difficult to work with.

k-grams are measures of frequency for strings of limited length. We define the *left k-gram* of a word to be the string that begins with the word and extends left, up to and including the first non-stopword. Similarly, the *right k-gram* of a word is the string that begins with the word and extends right, up to and including the first non-stopword. What constitutes a stopword might vary with application domain; we use the stopword list from Wordnet 2.1 in our work.

In our ordinary example sentence, the left k-gram of ‘attack’ is “expect that the attack” ($f = 50$) and the right k-gram is “attack will happen” ($f = 9260$). In the sentence with a substitution, the left k-gram of ‘campaign’ is “expect that the campaign” ($f = 77$) and the right k-gram is “campaign will happen” ($f = 132$).

We expect that, in general, k-grams will be smaller for sentences containing a substitution, although in the example this is only true for the right k-gram. Left and right k-grams capture significantly different information about the structure of sentences, which is not surprising given the linear way in which English is understood.

d) Hypernym Oddity (HO): The hypernym of a noun is a noun or noun phrase that is more general in a taxonomy of meaning. For example, the hypernym of ‘cat’ is ‘feline’.

Hypernyms themselves have hypernyms, so there are chains of increasing generality. For example, a chain contain ‘cat’ is: “kitty; house cat; cat; feline; carnivore; eutherian mammal”. Obviously the frequencies of the nouns and noun phrases along such a chain vary widely. A given word often has several hypernyms, usually reflecting different possible meanings.

Now consider the frequencies of a sentence (considered as a bag of words) and the same sentence where a particular word has been replaced by its hypernym. If the word is contextually appropriate then the sentence with the hypernym is likely to be more unusual, perhaps sounding a bit pompous (“the feline sat by the fire purring.”). On the other hand, if the word is not contextually appropriate, the sentence with the hypernym is likely to be more usual, since the hypernym is a more general concept. Hence we define the hypernym oddity of a sentence with respect to a particular word as:

$$\text{HO} \;=\; f_H - f$$

where f is the frequency of a sentence, regarded as a bag of words; and f_H is the frequency of a bag of words in which the word under consideration has been replaced by its hypernym. We expect this measure to be close to zero or negative when the word is contextually appropriate, but positive when the word is contextually inappropriate, and so probably a substitution.

For our example sentences, one hypernym for ‘attack’ is ‘operation’ and one hypernym for ‘campaign’ is ‘race’. The relevant frequencies for these sentences are given in Table II, giving hypernym oddity scores of $-1.11M$ for the ordinary sentence and $340,000$ for the sentence containing a substitution.

Because a given word typically has more than one hypernym, we can define the maximum, minimum, and average hypernym oddities over the possible choices of hypernym.

Hypernym chains tend to alternate between quite ordinary words and quite technical words. Hence the use of hyponyms (words below the word of interest in such a chain) give qualitatively similar results.

TABLE II
HYPERNYM EXAMPLES

Bag of words	Frequency
we expect that the attack will happen tonight	$f = 2.42M$
we expect that the operation will happen tonight	$f_H = 1.31M$
we expect that the campaign will happen tonight	$f = 1.63M$
we expect that the race will happen tonight	$f_H = 1.97M$

e) Pointwise Mutual Information (PMI): Pointwise mutual information attempts to measure the strength of an association between a word and some other string, either a word, a phrase, a sentence, or an entire document. We adapt this idea to measure the strength of association between a word that may be a substitution, and phrases of increasing length adjacent to it.

Consider a word of interest and an adjacent region of the sentence. The pointwise mutual information of the pair is given by:

$$\text{PMI} \;=\; \frac{p(\text{word})p(\text{adjacent region})}{p(\text{word + adjacent region})}$$

where $p()$ is a probability, and $+$ is concatenation in either direction, that is of the word with a phrase that follows it, or of the word with a phrase that precedes it. We can approximate the required probabilities as inverse frequencies. This results in values that are extremely small, so we take the reciprocal and define:

$$\text{PMI} \;=\; \frac{f(\text{word})f(\text{adjacent region})}{f(\text{word + adjacent region})}$$

where the frequencies are for quoted string searches. With this definition, PMI values are larger for words that are more unusual in their context. These values are so large that we present them divided by 10^9 to make them more readable.

The pointwise mutual information measure (PMI) is used extensively in data mining, and was introduced into text mining by Turney [14]. The advantage of using PMI for substitution detection is that it goes beyond our k-grams and sentence bag of words measures since it uses the frequency information of the constituents of phrases or sentences. The intuition behind applying the PMI measure is that if the target word is not contextually inappropriate (not substituted), then it should be a part of some stable phrase. Such a stable phrase should occur on the Web (or a suitably large corpus) more often than random chance dictates.

Although the PMI formula uses probabilities of occurrences and co-occurrence, it has been commonly approximated by ratios of numbers of occurrences on the Web (or any sufficiently large repository) [14]. While no one has formally studied the accuracy of such an approximation, we can intuitively justify it by assuming that all pages are of approximately the same size, and the evaluated words are distributed uniformly throughout each page.

We calculate a family of pointwise mutual information measures using nested adjacent regions that increase in length until their observed frequencies drop to zero. Adjacent regions can precede or follow the word of interest. We compute the maximum pointwise mutual informations over all choices of adjacent regions of text.

PMI scores for some of the adjacent regions of the example sentence are shown in Table III; and for some of the adjacent regions in the sentence with a substitution are shown in Table IV.

The string “attack will happen tonight” occurs only once, and the string “campaign will happen tonight” never occurs, so these regions cannot be made larger.

Performing a full or partial grammatical parse and limiting the application of measures to the related components of sentences would probably increase the effectiveness of these measures. However, it would significantly slow down the processing.

IV. Experiments and Results

A. Frequency Oracles

We use the Yahoo web service search interface as a frequency oracle for frequencies of words, bags of words, and quoted strings. Yahoo claims to index about 20 billion pages.

There are several practical issues in using such frequency oracles. First, we use the number of pages returned by a search as a surrogate for the natural frequency of the search terms. This implicitly assumes that every word occurs the same number of times in each page where it appears. For sets of words, this assumption also fails to capture how far apart they appear in the page. We justify this on the grounds that most searches return a large number of pages, so considerable smoothing occurs.

Second, the way in which stopwords are handled by search engines is opaque. For example, at Yahoo, the quoted string “chase the dog” occurs 2990 times, while the quoted string “chase dog” occurs only 1290 times, so clearly stopwords are taken into account in such searches. However, the bag of words {chase the dog} occurs 7,610,000 times while the bag of words {chase dog} occurs only 7,030,000 times, which seems counterintuitive.

Third, word order seems to be significant, even in bag of word searches. The bag {natural language processing} occurs 5,750,000 times while {natural processing language} occurs 6,210,000 times.

Fourth, the frequencies reported by Yahoo and Google are substantially different in ways that cannot be easily accounted for on the basis of different numbers of kinds of pages indexed. For example, Google reports that the quoted string “chase the dog” occurs 18,200 times, a factor of 6 greater than Yahoo’s frequency.

These issues mean that frequency data must be treated with caution, especially when the frequencies are low. However, search engines index extremely large amounts of textual data, so we expect that, in a broad sense, they capture properties of natural language well.

We also use the British National Corpus (BNC) [15] as a resource for word frequencies. The BNC contains 100 million words collected from both spoken and written English. Frequency ranked lists of words, including lists for particular parts of speech, are derived from the corpus.

B. Test Data

We apply our measures to two datasets, one derived from the Enron email corpus, and the other from the Brown news corpus.

The Enron email corpus was made public as the result of the prosecutions of Enron personnel. It contains slightly fewer than half a million emails (many of them duplicates) to and from Enron staff over a period of three and a half years. The authors of the emails never expected that it would be made public, so it is a good sample of informal writing, by a large number of authors, from many backgrounds. As such, it is a good surrogate for the kinds of messages that might be intercepted in an intelligence or law-enforcement context.

TABLE III

PMI SCORES FOR THE EXAMPLE SENTENCE

word = ‘attack’, f(attack) $= 174M$			PMI
‘the attack’	f(the attack) $= 19.1M$	f(the) $= 6,580M$	59.94
‘that the attack’	f(that the attack) $= 703,000$	f(that the) $= 943M$	233.4
‘expect that the attack’	f(expect that the attack) $= 50$	f(expect that the) $= 2.24M$	7795
‘attack will’	f(attack will) $= 811,000$	f(will) $= 2.67B$	572.8
‘attack will happen’	f(attack will happen) $= 9260$	f(will happen) $= 19.3M$	362.7

TABLE IV

PMI SCORES FOR THE SENTENCE WITH A SUBSTITUTION

word = ‘campaign’, f(campaign) $= 167M$			PMI
‘the campaign’	f(the campaign) $= 22.8M$	f(the) $= 6,580M$	48.20
‘that the campaign’	f(that the campaign) $= 575,000$	f(that the) $= 943M$	273.8
‘expect that the campaign’	f(expect that the campaign) $= 77$	f(expect that the) $= 2.24M$	4858
‘campaign will’	f(campaign will) $= 2.43M$	f(will) $= 2.67B$	183
‘campaign will happen’	f(campaign will happen) $= 132$	f(will happen) $= 19.3M$	24417

For the Enron email corpus, sentences containing between five and fifteen words, inclusive, were selected, giving a total of 712,662 candidate sentences.

We limited our attention to substitution of nouns, since these carry the greater part of the content of sentences. Sentences were uniformly randomly selected from the set of candidate sentences, but discarded if the first noun in the sentence (a) was not present on the BNC noun list, or (b) did not have a hypernym known to Wordnet. This removed primarily sentences that were not English.

A set of 1714 sentences representative of informal written English resulted. A new set of 1714 sentences, each containing a substitution, was constructed from the Enron set by replacing the first noun in each sentence by the noun with next-highest frequency on the BNC noun frequency list. Some examples of pairs of ordinary sentences and sentences with a substitution are:

an agent will assist you with checked baggage
an vote will assist you with checked baggage
my lunch contained white tuna she ordered a parfait
my package contained white tuna she ordered a parfait
please let me know if you have this information
please let me know if you have this men

We therefore have two sets of sentences, labelled as ordinary or containing a substitution, to which we can apply our measures.

The Brown news corpus contains about one million words, from a variety of more formal texts, including news and commentary. We chose this set for comparison because we expect that the writing style is much more formal than in the Enron emails, all the more so as it was collected in 1961. We expected that substitutions might be easier to detect in this data.

The same processing was carried out to select a set of 566 ordinary sentences; and the first noun in each sentence was replaced using the same algorithm to produce a set of 566 sentences that contained a substitution. Some examples of pairs of ordinary sentences and sentences with a substitution are:

it was one of a series of recommendations by the texas research league
it was one of a bank of recommendations by the texas research league
the remainder of the college requirement would be in general subjects
the attendance of the college requirement would be in general subjects
a copy was released to the press
a object was released to the press

These sentences are indeed more formal than those of the Enron corpus.

TABLE V

Optimal boundary values based on information gain

Measure	Boundary
Sentence oddity	4.6
Enhanced sentence oddity	0.98
Left k-gram	155
Right k-gram	612
Average k-gram	6173
Minimum hypernym oddity	-89129
Maximum hypernym oddity	-6
Average hypernym oddity	-6
Maximum PMI	1.34

TABLE VI

Sentence oddity performance

Corpus	Detection Rate (%) split (10-fold)	False Positive Rate (%) split (10-fold)	Area under the ROC curve
Enron	51 (57)	21 (25)	0.6672
Brown	30 (65)	15 (43)	0.6219

TABLE VII

Enhanced sentence oddity performance

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	72 (73)	23 (23)	0.7744
Brown	59 (63)	17 (18)	0.7576

C. Experiments

An appropriate decision surface for each measure was determined by training it using the J48 decision tree provided by Weka (`www.cs.waikato.ac.nz/ml/weka`) with the default parameters. At the same time, this gives an estimate of the performance of each measure as an independent classifier. Performance was estimated using a 75% training set and 25% test set, and also by using 10-fold cross validation.

Since each of decision trees for individual measures is training on a dataset with a single attribute (the measure score), the attribute decision boundary at the root of each tree represents the best split point based on information gain. These values provide an insight into the meaningful distinctions for each measure. Boundary values are shown in Table V. For example, these boundaries suggest that a sentence contains a substitution if its sentence oddity is greater than 4.6, that is if the frequency of the sentence words without the target is more than 4.6 times the frequency of the sentence including the target word. Similarly, a sentence contains a substitution if its left k-gram occurs fewer than 155 times or its right k-gram occurs fewer than 612 times.

D. Performance of Individual Measures

For the Enron corpus, we use a set of 1714 ordinary sentences and 1714 sentences containing substitutions. For the Brown corpus, we use a set of 566 ordinary sentences and 566 sentences containing substitutions. The set of sentences was generated incrementally, so we were able to observe the measure values as the number of sentence increased. These values were remarkably stable with respect to the size of the datasets once the number of sentences exceeded 200, suggesting that they would not change significantly if dataset sizes were increased further.

Tables VI–XIV show the detection rate, that is the percentage of sentences containing a substitution that is detected, and the false positive rate, that is the percentage of ordinary sentences that are classified as containing a substitution. Two values are provided for each rate: the first is the rate for a 75%-25% training/test set split; the second is the rate for 10-fold cross-validation.

We also show the area under the ROC curve for the class of sentences containing a substitution. This gives a sense of how well each measure is performing with respect to the trade-off between false rejection rate and false acceptance rate.

The sentence oddity measure is mediocre both at detecting sentences containing substitutions and detecting ordinary sentences, and noticeably worse for the Brown corpus than for the Enron corpus. Enhanced sentence oddity is a little better at detecting sentences containing substitutions for both datasets.

The left k-gram is another weak measure for both corpora, but the right k-gram has a significant detection rate for the Enron corpus. The average k-gram mixes these two measures, but produces little improvement.

All three hypernym measures are quite weak for

TABLE VIII

LEFT K-GRAM PERFORMANCE

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	56 (53)	33 (25)	0.6403
Brown	40 (39)	26 (26)	0.5981

TABLE IX

RIGHT K-GRAM PERFORMANCE

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	84 (81)	52 (47)	0.6791
Brown	27 (41)	9 (14)	0.6360

TABLE X

AVERAGE K-GRAM PERFORMANCE

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	56 (56)	25 (21)	0.6768
Brown	23 (50)	10 (29)	0.6237

TABLE XI

MINIMUM HYPERNYM ODDITY PERFORMANCE

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	66 (45)	52 (33)	0.5735
Brown	0 (43)	0 (41)	0.5522

TABLE XII

MAXIMUM HYPERNYM ODDITY PERFORMANCE

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	57 (55)	30 (29)	0.6330
Brown	0 (52)	0 (45)	0.5627

TABLE XIII

AVERAGE HYPERNYM ODDITY PERFORMANCE

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	43 (42)	21 (21)	0.6068
Brown	42 (40)	20 (25)	0.5742

TABLE XIV

MAXIMUM PMI PERFORMANCE

Corpus	Detection Rate	False Positive Rate	Area under the ROC curve
Enron	49 (54)	24 (23)	0.7064
Brown	99 (67)	69 (43)	0.6989

the Enron corpus. For the Brown corpus and the 75%–25% split, both minimum and maximum hypernyms predict every sentence to be ordinary. We suspect that this is because much of the content of the Brown corpus is formal writing. In this setting, hypernyms of nouns tend to be extremely formal or technical. As a result, f_H tends to be small, making the hypernym scores large and negative (and so predicting ordinary sentences).

The maximum PMI is quite a weak measure on the Enron corpus, but detects sentences with substitutions very strongly on the Enron corpus with the 75%–25% split. Again, we suspect this is a consequence of the relatively formal sentence structure.

Overall the results are less predictive for the Brown corpus, but in interesting ways. The Brown corpus was collected in 1961, and captures primarily formal writing. It is possible that, with the passage of time and a general loosening of the rules of grammar, today's text repositories do not represent co-occurrence frequencies well for such data. It may also be that phrases have remained in use over this time period, so that measures such as PMI that access deeper structures in sentences are less affected.

Since our primary objective was not to compare existing measures but rather to suggest our own measures to tackle this novel task, we did not compute the statistical significance of the performance of the measures we have suggested. Given the stability of the results with respect to the number of sentences and the unlimited availability of the data for testing, we are confident that any necessary level of statistical significance can be established by following the same methodology.

E. Combining Predictors

Most of the measures described in the previous subsection perform poorly at detecting sentences with substitutions, or detecting ordinary sentences,

TABLE XV
OVERALL PERFORMANCE: DECISION TREE

Corpus	Detection Rate (%) split (10-fold)	False Positive Rate (%) split (10-fold)	Area under the ROC curve
Enron	95 (94)	11 (11)	0.9844
Brown	84 (91)	16 (15)	0.9838

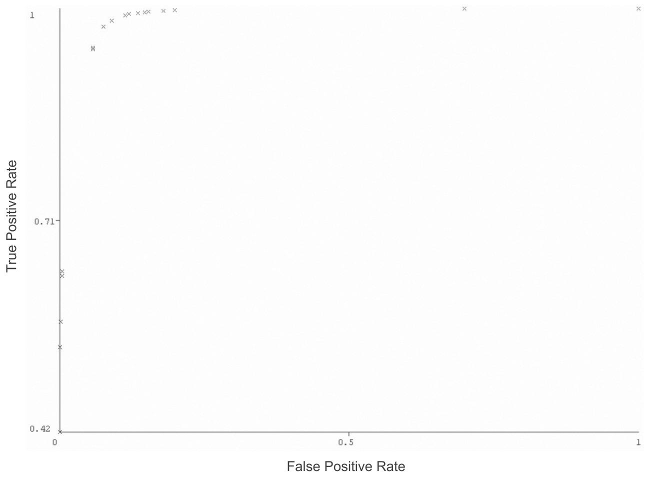

Fig. 1. ROC curve for Enron corpus combined predictor

or both. Fortunately, these measures look for different sentence properties, so they make errors on different sentences. As a result, when the measures are combined, prediction accuracy is very high.

We consider two different methods of combining the individual measures. A decision tree using all of the measure values as attributes applies information gain criteria to selecting the measures and applying them in an effective order. The performance of decision trees trained on all of the measures is shown in Table XV. In the Enron corpus, this combination of measures can detect sentences containing a substitution with accuracies in the mid-ninety percent range, with a false positive rate of around ten percent. Performance is worse for the Brown corpus.

Figures 1 and 2 show ROC curves for the Enron and Brown corpora respectively. The sharp knee in the upper left hand corner of each shows how well these combined measures perform.

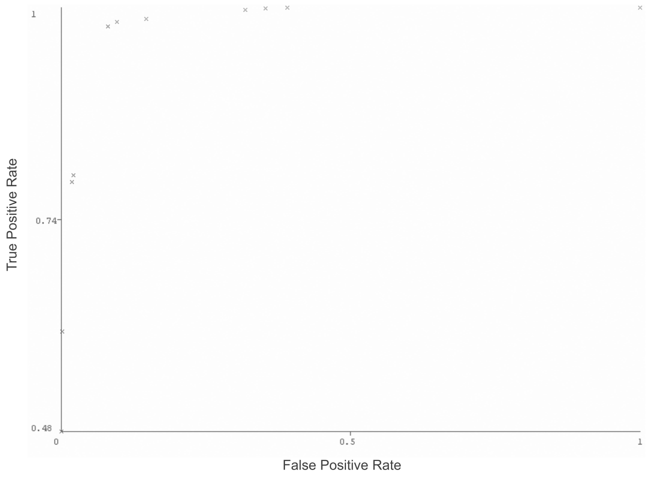

Fig. 2. ROC curve for Brown corpus combined predictor

Our second method of combining measures is to use a random forest [16]. Random forests grow large numbers of unpruned decision trees, using samples drawn from the original data with replacement, until the full size of the training data is reached. Typically about one-third of the objects are never selected, and these act as an immediate test set. A fixed number of candidate attributes are chosen afresh for the construction of each tree node, with the best being chosen in the usual way. This method of construction prevents random forests from overfitting the data, and ensures that important attributes play an important role, while those with little predictive power do not.

A random forest with 50 trees, and a branch size (i.e. Mtry) of 4 was trained on a dataset of all of the measure values, using a 75%-25% split. Its performances are shown in Table XVI. For the Enron corpus, combining the measures using a random forest has a detection rate in the low ninety percents, with a false positive rate around ten percent, comparable to the combined decision tree. Again, the results for the Brown corpus are worse.

F. Discussion

By manually inspecting the first 100 sentences from the Enron corpus, we discovered that there were three approximately equally contributing sources of classification errors:

1) The original sentences were either not grammatically correct or too short, and so seemed

TABLE XVI
OVERALL PERFORMANCE: RANDOM FOREST

Corpus	Detection Rate (%) split	False Positive Rate (%) split
Enron	90	11
Brown	83	13

unusual even before substitution, for example "body and the other in jpg".

2) The substituted word was the only word in the sentence that was not a stopword or some other very frequent word. Thus there was no visible change in the values of the measures, for example "investigation me if you need any more input".
3) The substitution was, by coincidence, not contextually unusual.

All of the sources of errors seem to result from properties of the test data used and the way in which we create sentences with substitutions, rather than from the weaknesses in the measures used. This suggest that we have achieved approximately the upper bound of performance for the family of measures that we studied and the datasets that we used.

Detecting substitutions is a difficult problem, so it is not surprising that individual measures perform quite poorly. However, different measures look for different kinds of contextual discontinuities and so make errors in different sentences. When the individual measure scores are combined in an ensemble, the overall performance is much better than that of any individual measure.

We did not include measures based on Hidden Markov Models, which are popular in speech recognition, for two reasons. First, PMI measures can be considered as generalizations of Markov models since they look at co-occurrences on both sides, rather than only on the left side as speech recognition models do. Second, tests with Markov models on the Enron corpus indicated than they performed approximately 20-30% worse than PMI-based measures. We are leaving for future research more detailed comparison and incorporation of HMMs into a combined classification model.

We also ran almost all of the above tests using MSN as the frequency oracle and obtained essentially the same results. This indicates that, although discrepancies between the search engines may exist, the choice of oracle does not affect the performance of the measures based on phrase frequencies. This is not surprising since the measures are based on ratios, and the task itself is noisy. This finding is consistent with similar findings that use frequency information for other text-mining tasks, for example, for fact seeking [17]. The finding also suggests that combining several oracles may improve the results slightly, but we leave that for future research.

The false positive rate is the performance-limiting component of such ensembles, since the overwhelming majority of sentences in real applications will not contain substitutions. For example, if 1 sentence in a million contains a substitution, then a false positive rate of 10% selects 100,000 innocent sentences as well as (almost certainly) the one suspicious sentence.

However, even with a significant false positive rate, the application of an ensemble of measures acts as a filter to reduce the fraction of messages that need to be considered by subsequent analysis. Furthermore, the words(s) suspected of being substitutions can be labelled by the ensemble, making subsequent examination even easier. For example, words suspected of being substitutions could be color-coded to make it easier and faster for a human analyst to decide whether or not they were suspicious.

There is a further opportunity to compensate for weaknesses in the ensemble performance by considering how classification of sentences is used to determine classification of entire messages. For example, if we use the criterion that a message is suspicious if it contains one sentence with a substitution, and we assume that the average message length is ten sentences, then a false positive rate of 10% selects almost every message as suspicious. On the other hand, with a detection rate of 90%, the possibility of missing a message that does contain a substitution is vanishingly small: 10^{-10}. If the criterion that a message is suspicious is weakened to requiring that it contains at least three sentences containing a substitution (which is quite reasonable in practice since a message is about a topic, and this topic may be the thing that must be concealed), then the false positive rate per message drops to around 1.2%. This reduces the amount of text to be processed further by more than a factor of 8.

We have experimented with replacing words by new words of significantly different frequency. In preliminary results, replacing a noun with a noun whose rank is half that of the word it replaces produces performances similar to those reported here. Replacing a noun by a noun whose rank is twice that of the noun it replaces appears to make detecting substitutions a little easier. We expect that this is at least partly because the replacement word is automatically rarer, and so inherently more unusual in any context. We are continuing to explore these issues.

Since our objective was to explore the feasibility of this approach, we were not concerned with real-time response. The complete run of our test set took many hours for each individual measure. The bottleneck is querying the underlying search engines, which was necessarily parallelized on multiple systems. Runtime would be much improved with direct access to the search engine.

V. Conclusions

The problem of detecting when word substitution has occurred has a role to play in settings such as counterterrorism and law enforcement, where large amounts of message traffic may be intercepted in an automated way, and it is desirable to reduce the number of messages to which further analysis must be applied. The existence of simple mechanisms such as watchlists of significant words may actually make the discovery of illicit groups easier, because they must react to the existence of watchlists while innocent groups are either unaware of them, or do not alter their messages.

If the goal of illicit groups is to evade automated detection, then it is important that the word substitutions should look as normal as possible from a syntactic perspective (whereas if humans were searching for suspicious messages, a much more semantic form of substitution would be required).

We have addressed the problem of detecting the kind of substitutions that might be made in response to watchlist scanning: replacing words with words of similar frequency. Such substitutions are quite effective in obfuscating content, as demonstrated by the low detection rates and high positive rates of most of the measures we have designed. However, these families of measures make errors on different sentences so that, when they are combined, the overall detection rates are close to 90% or better and the false positive rates fall to around 10%. These rates make the combined predictors usable at the scale of messages.

English is extremely variable, so that there are examples of extremely unusual sentences, especially in the Enron email corpus. So, in a sense, it is not surprising that it is so difficult to detect abnormal combinations of words caused by substitution, since many ordinary sentences also contain abnormal combinations. This variability also means that many ordinary sentences, and indeed fragments of sentences, are not captured by search engines, so we are unable to estimate their frequencies. This also limits the performance of the measures.

Although the Brown corpus contains more formal writing, it is not easier to detect substitutions in this setting. This is perhaps surprising, but may reflect changes in language patterns, properties of formal writing or both.

Although this work has been based on English sentences, there seems no strong reason why the results should not extend to other languages, especially uninflected languages where word order is important. Frequency oracles for other languages are available, but are based on much smaller samples of texts.

A number of limitation have been mentioned throughout the paper, and we are planning to address them in future research. Some specific issues are: testing substitutions of verbs rather than nouns, analyzing groups of messages instead of single ones, applying parsers to find important relations between words that we now approximate using k-grams, testing multiple substitutions within a single sentence (for example, “The alcohol is in the bar” instead of “the bomb is in position”), using a wider variety of test sets, and testing specific known ‘red flag’ terms. It will be also interesting to investigate how correlation between substitutions can be exploited to increase accuracy, and perhaps even to guess what original words were replaced.

We believe that measures to detect substitution can be used in other applications involving automated text analysis, such as deception detection, authorship identification, data de-identification, psychiatry, and analysis of financial reports.

REFERENCES

[1] D. Skillicorn, "Beyond keyword filtering for message and conversation detection," in *IEEE International Conference on Intelligence and Security Informatics (ISI2005)*. Springer-Verlag Lecture Notes in Computer Science LNCS 3495, May 2005, pp. 231–243.

[2] "Message on MySpace prompts school to beef up security," www.10news.com/news/9150360/detail.html, 2006.

[3] D. Roussinov and J. Roubles, "Applying question answering technology to locating malevolent online content," *Decision Support Systems*, to appear.

[4] P. Brown, P. deSouza, R. Mercer, V. D. Pietra, and J. Lai, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[5] S. Fong, D. Skillicorn, and D. Roussinov, "Measures to detect word substitution in intercepted communication," in *Intelligence and Security Informatics, IEEE International Conference on Intelligence and Security Informatics, ISI 2006, San Diego, CA, USA, May 23-24*, ser. LNCS 3975, 2006.

[6] J. Bilmes and K. Kirchhoff, "Factored language models and generalized parallel backoff," in *Proceedings of HLT/NACCL*, 2003.

[7] A. R. Golding and D. Roth, "A Winnow-based approach to context-sensitive spelling correction," *Machine Learning, Special issue on Machine Learning and Natural Language*, 1999.

[8] H. Lee and A. Ng, "Spam deobfuscation using a Hidden Markov Model," in *Proceedings of the Second Conference on Email and Anti-Spam*, 2005.

[9] D. Roussinov, L. Zhao, and W. Fan, "Mining context specific similarity relationships using the World Wide Web," in *Proceedings of the 2005 Conference on Human Language Technologies*, 2005.

[10] D. Roussinov and L. Zhao, "Automatic discovery of similarity relationships through web mining," *Decision Support Systems*, pp. 149–166, 2003.

[11] K. Olsen and J. Williams, "Spelling and grammar checking using the Web as a text repository," *Journal of the American Society for Information Science and Technology*, vol. 5, no. 11, pp. 1020–1023, 2004.

[12] R. F. i Cancho and R. Solé, "The small world of human language," *Proceedings of the Royal Society of London Series B – Biological Sciences*, pp. 2261–2265, 2001.

[13] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, 2001.*, 2001, pp. 533–536.

[14] P. Turney, "Mining the web for synonyms: PMI-IR versus LSA on TOEFL," in *Proceedings of the Twelfth European Conference on Machine Learning (ECML-2001)*, 2001, pp. 491–502.

[15] "British National Corpus (BNC)," 2004, **www.natcorp.ox.ac.uk**.

[16] L. Breiman, "Random forests–random features," Department of Statistics, University of California, Berkeley, Tech. Rep. 567, September 1999.

[17] S. Dumais, M. Banko, E. Brill, J. Lin, and A. Ng, "Web question answering: Is more always better?" in *Proceedings of the 25th Annual international ACM SIGIR conference on Research and Development in Information Retrieval*, 2002, pp. 11–15.