Detection of Word Substitution In Intercepted Communication

S . Venkata Lakshmi,
Assistant Professor
CSE Dept,KMMITS,
Tirupati

J . Pushpa Kumari,
Assistant Professor
CSE Dept,KMMITS,
Tirupati

P . S .Hemalatha,
Assistant Professor
CSE Dept,KMMITS,
Tirupati

Abstract:- **When messages may be intercepted because they contain certain words, terrorists and criminals may replace such words by other words or locations. If the replacement words have different frequencies from the original words, techniques to detect the substitution are known. We address the problem of discovering such substitutions when the original and substitute words have the same natural frequency. Each of these measures individually is a weak detector. However, we show that combining them produces a detector that is reasonably effective.**

Keywords: k-gram, frequencies, hypernym oddity, sentence oddity, BNC corpus

1. INTRODUCTION

Terrorists and criminals must be aware of the possibility of interception whenever they communicate by phone or email. In particular, terrorists must be aware of systems such as Echelon that examine a very large number of messages and select some for further analysis based on a watch list of significant words.

Given that it may not be possible to evade some examination of their messages, terrorists and criminals have two defensive strategies: encryption and obfuscation. The problems with encryption are that it draws immediate attention to messages and so permits at least meta-analysis; and it may be that there are backdoors to commonly available encryption methods. Obfuscation tries to hide messages in the background of the vast number of other messages, replacing words that might trigger attention by other innocent-sounding words or locutions. For example, al Qaeda, for a time, used the word `wedding' to mean `attack'.

One way to conceal content is to encrypt the messages, but this strategy has a number of drawbacks. First, encryption draws attention to messages, making techniques such as traffic analysis easier to apply. Second, encryption is hard to use with readily available components in some settings, for example cell phone calls. Third, it is hard to be sure exactly how robust encryption is in practice, since agencies such as the U.S. NSA do not reveal their decryption capabilities and there are persistent rumors of back doors into common encryption systems. messages is to replace significant words with other words or locutions that are judged less likely to attract attention. For example, it is known that Echelon scans for a list of significant words or phrases, and terrorists would presumably wish not to use these words in their messages. The difficulty is that, while it is clear that some words must be on these lists (e.g. `nuclear'), it is difficult to guess how long such lists are (`fertilizer'?). Replacing words with more innocuous words in real time, for example during a cell phone call, is not easy, and it is likely that the replacement words will differ in obvious ways from the words they replace. For example, humans do not appear to have an intrinsic understanding of word frequencies, so it is likely that a word and its replacement would have significantly different frequencies However, replacement of words by words of similar frequency becomes possible given access to a word-frequency table.

Consider the sentence the bomb is in position". A word of similar frequency to `bomb' is `alcohol'. A human might well consider the sentence the alcohol is in position" to be slightly odd, but on the basis of semantic information about the typical uses of alcohol. We are interested in whether such substitutions can be detected using semantic information only indirectly via, for example, word and word-group frequencies. Consider the sentence the attack will be tomorrow". Using the al Qaeda substitution, we get the wedding will be tomorrow" which is designedly a natural sounding sentence. However, `attack' is the 1072nd most common English word according to the site www.wordcount.org/main.php, while `wedding' is the 2912th most common, so the substantial frequency difference might make this substitution detectable using the approach described above. On the other hand, if the word `attack' is replaced by the word `complex' which has similar frequency, than any human will be able to detect that the sentence \the complex will be tomorrow" is extremely unusual. However, detecting this kind of substitution automatically using software has not been attempted, except in a very preliminary way.

2. RELATED WORK

Detecting words out of context can also be used to detect (and correct) misspellings This problem differs from the problem addressed here because the misspelled words

are nonsense, and often nonsense predictably transformed from the correctly spelled word, for example by letter reversal.

Detecting words out of context has also been applied to the problem of spam detection. For example, Spam Assassin 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 detect word-level manipulations typical of spam using Hidden Markov Models. As part of this work, they address the question of whether an email contains examples of obfuscation at all. They expected this to be simpler than the problem they set out to address {recovering the text that had been obfuscated {but remark that detecting obfuscation at all is `surprisingly difficult' 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.

3.STRATEGIES FOR DETECTING SUBSTITUTION

We consider three ways in which a word may appear unusual in a particular context. All depend on an intuition that the substituted word appears unusual in context because its semantics does not with the semantics of the context. Substitution is purely syntactic, based on single word frequencies, but its effects are semantic and so potentially detectable. Here are three measures that may reveal this form of discrepancy:

1. When a word substitution has occurred, the frequencies of pairs of a given word with its neighbors on either side may decrease because the word is not as appropriate in these context as the original word it replaces would have been. This intuition extends to larger contexts, such as all of the n-grams containing the substituted word.
2. When a word substitution has occurred, the sentence should be of low frequency, since the substituted word presumably does not occur often in such a context. Hence we compute the ratio of the frequency of the sentence, with the substituted word omitted, and considered as a bag of words, to the frequency of the entire sentence, again as a bag of words. A sentence containing a substituted word should produce a large ratio using this measure.
3. If a noun is appropriate in its context, then replacing it by its hypernym2 should also produce a meaningful sentence. For an ordinary sentence, the replacement by a hypernym tends to produce a more unusual sentence, and hence a reduced frequency. For a sentence containing a substituted word, replacement by a hypernym tends to produce a more common sentence because the hypernym is a more general word and so may occur more often.

3.1 Extracting a sentence dataset

We use the Enron email dataset as a source of sentences. The Enron email dataset contains emails sent and received by Enron employees in the three and a half years before the collapse of the company. These emails are informal documents, that received little or no editing at the time, and which their senders did not expect to be made public. They are therefore good representations of what intercepted communications might look like. It will become clear from the results that the use of such real data is important { some of the problems encountered are the result of informal sentence structures that would not have been present in more artificial data.

Since Enron emails contain many strings that are not English words, for example words in
other languages and strings such as acronyms, we use the British National Corpus (BNC) to discard any string that appears not to be an English word, and also as a canonical source of frequencies of English words.

Sentences containing substitutions were constructed from this set by binding the first noun that did not appear in a stop word list, and replacing it by the next most frequent noun from the BNC corpus. The stop word list in WorldNet 2.0 was used. A random sample of 200 sentences was drawn from this set. The size of this sample is constrained by the time taken to process the set. Only sentences for which a hypernym exists in WorldNet for the substituted word were retained,reducing the set of 200 sentences to a set of 98 ordinary sentences and 98 sentences containing a substitution which are used throughout the paper.

3.2 Frequencies

Frequencies of sets of words are measured by using the API at Google. There are a number of
aspects of the way frequencies are computed by Google that complicate its use as a frequency oracle. First, the frequencies returned via the API and via the web interface are substantially different; for consistency we use the API frequency values throughout, but these differences suggest some uncertainty. Second, the Google index is updated every 10 days or so, but this is not trivially detectable, so frequencies may be counted from different instantiations of the index (large frequencies are rounded so this makes little difference, except for rare strings). Third, the way Google handles stop words is not transparent, and makes it impossible to invoke exactly the searches we might have wished. For exam ple, \chase the dog" occurs 9,580 times whereas \chase dog" occurs 709 times, so quoted string searches clearly do not ignore stop words.

On the other hand, the bag of words search *f*chase the dog*g* occurs 6,510,000 times while *f*chase dog*g* occurs only 6,490,000 times, which seems counterintuitive. Fourth, the order of words is significant, even in bag of word searches. Frequencies returned by Google should be adjusted to redirect the fact that the strings indexed by Google are a sample of the universe of English strings in use. We ignore this issue on the grounds that Google

provides a very *large* sample, but sampling artifacts are occasionally visible in the results.

3.3 k-gram measures

When a substitution has occurred, we expect that the frequencies of n-grams that contain the substituted word will be lower than expected; in other words, a sliding window of size *n* should show a decrease in frequency whenever it contains the substituted word. However, the structure of the Google API interface makes it difficult to count the frequencies of n-grams as such. Instead we measure the frequency of a generalized n-gram which we call a k-gram. The k-gram of a substituted word is the string containing that word and its context up to and including the first non-stop word to its left, and the first nonstop word to its right. For example, ten miles is a long way to walk", the k-gram for `miles' is ten miles is a long", and the k-gram for `way' is long way to walk". The frequency of the resulting exact string is determined from Google.

A threshold for determining when a word is a substitution was learned using a decision tree
whose only attribute is the measure values for the two classes: the k-grams of the original set of 98 sentences and the k-grams of the 98 sentences with substitution. The decision boundary based on this model is 4, that is any k-gram whose Google frequency is at least 4 can be considered as coming from an ordinary sentence.

3.4 Sentence oddity

Sentence oddity measures are designed to measure the frequency of an entire sentence. Because most sentences do not appear verbatim even once in a large text repository, obtaining such frequencies comes at the expense of ignoring the order of the sentence words. In general, if a word is discarded from a bag of words, the frequency of the smaller bag should be greater than that of the original bag. However, if the bag of words was a sentence with the word order ignored, and the discarded word was meaningful in the context of the sentence, then we might expect that the difference in frequency might be moderate. If the discarded word was not meaningful in the context of the sentence, then the difference in frequency might be much greater. Hence we deferent sentence oddity as:

$$_{sentence\ oddity} = \textit{frequency of bag of words with word discarded/ frequency of entire bag of words.}$$

The more unusual the discarded word was in the context of its sentence, the greater we expect the sentence oddity to be.

3.5 Semantic oddity

If a word is a substitution, then we expect that word not to first into the context well. If the substituted word is, in turn, replaced by a related word, the frequency of the resulting sentence will change, and this change will reflect something about how unusual the original substitution was. This requires a way to find related words, which is fundamentally a semantic issue, but there are sources of such words, for example WorldNet. The hypernym of a noun is the word immediately above it in the ordinary ontology of meanings; for example, the hypernym of `car' is `motor vehicle'.

We had expected that, when a normal word is replaced by its hypernym, the frequency of the resulting sentence would stay the same or increase; while when a substituted word is replaced by its hypernym the frequency of the resulting sentence would decrease. In fact, the chain of hypernyms for many words exhibits an oscillating structure, moving from technical terms to common terms and then back to technical terms, and so on. For example, a chain containing `attack" is (from the bottom): foray, incursion, attack, operation, activity, act, event" in which `attack' and `act' are simpler words than the others. Another chain is comprehension, understanding, knowing, higher cognitive process, process, cognition", in which `understanding', `knowing, and `process' are ordinary words while the other words in the chain are more technical.

In ordinary informal text, the nouns in use are likely to be close to the appropriate class words {using nonclass words tends to sound pompous. Substitution by a hypernym is likely to produce a more technical sentence, with a lower frequency. If the noun under consideration is already a substitution, however, it is less likely to be a simple word. Substitution by a hypernym may produce a less technical sentence with a greater frequency. The chain containing `complex' is: hybrid, complex, whole, concept, idea, mental object". In our example sentence, \the complex is tomorrow", replacement produces the whole is tomorrow" which is a much more common bag of words.We de¯ne the *hypernym oddity* to be:

$$\text{hypernym oddity} = fH \ ¡ \ f$$

Where *f* is the frequency of a sentence, regarded as a bag of words; and *fH* is the frequency of a bag of words in which the noun under consideration has been replaced by its hypernym. We expect this measure to be close to zero or negative for ordinary sentences, but positive for sentences that contain a substitution . Ordinary sentences, but positive for sentences that contain a substitution.

These three strategies, looking for frequencies of exact substrings of the sentence under consideration, looking for changes in frequency between the entire sentence and the sentence without the word under consideration, and looking for changes in frequency when the word under consideration is replaced by its hypernym (or other related words) can all suggest when a substitution has occurred. In the next section, we describe the exact measures we have used in our experiments.

4 . TECHNIQUES

4.1 Usable frequency data

In order to be able to measure the frequencies of sentences, sentence fragments, and bags of words, we must use data about some repository of text. The choice of repository makes a great deal of difference, since the better the match between the repository and the style of text in which substitutions may have occurred, the more accurate the prediction of substitutions will be. It is well known, for example, that perplexity, which measures a one-sided 2-gram frequency, is considerably reduced in sets of documents from a particular domain. We use Google as the source of frequency data, on the grounds that it indexes a very large number of English documents, and so provides a good picture of frequencies of English text. That said, it is surprising how often an apparently ordinary phrase occurs zero times in Google's document collection. *f*natural processing language *g* consistently produce different frequencies.

We use the number of pages returned by Google as a surrogate for word frequency. This fails to take into account intra word frequencies within each individual document. It also fails to take into account whether two words appear, say, adjacently or at opposite ends of a given returned document, which we might expect to be relevant information about their relationship. We have experimented with using locality information of this kind, but it does not improve performance.

4.2 Usable semantic data

The only semantic information we use is the hypernyms of nouns being considered. We get this information from WorldNet (wordnet.princeton.edu). In general, a word can have several hypernyms, so we collect the entire set and use them as described below. For example, the direct hypernyms of `complex' are `whole', `compound', `feeling', and `structure', derived from the di®erent meanings of `complex'.

4.3 Experimental data

In order to evaluate measures to detect substitutions, we need sets of reasonable sentences to use as data. Standard grammatical sentences, for example from news articles, do not make good test data because the kinds of sentences intercepted from email and (even more so) from speech will not necessarily be complete or formal grammatical sentences. A large set of emails was made public as the result of the prosecution of the Enron corporation. This set of emails was collected over three and a half years and contains emails from and to a large set of individuals who never imagined that they would be made public. This set of emails is therefore a good surrogate for the kinds of texts that might be collected by systems such as Echelon, and we use it as a source of informal, and so realistic, sentences.

The original set of sentences is useful because it lets us measure the false positive rate of the various measures. Also using a set in which the only difference is the occurrence of a substitution guarantees that performance differences do not arise from other features of the sentences.

5. RESULTS AND DISCUSSION

5.1 Experiments

We compute each of these measures for the set of 98 sentences containing substituted words, obtaining frequency data via the Google API. We also compute the measures for the original set of sentences without substitution as a way of assessing the false positive rates that each measure might generate. In a deployed system, the original sentence would not, of course, be available.The sample size is too small to estimate the robustness of the results, but we have preliminary results on a much larger sentence set which are consistent with those presented here.

5.2 k-grams

Recall that a decision boundary of 4 for the k-gram measure was estimated, based on the difference between the original and substituted sentence datasets. The prediction accuracy for sentences with substitutions was 81% using this boundary, but at the expense of a 47% false positive rate for the ordinary sentences. There are several reasons why the false positive rate on ordinary sentences is so high. First, some of the k-grams are quite long (8-12 words) so that the probability of any occurrences is inherently low (for example, \curious whether his rant was getting any traction"). Second, these k-grams often capture unusual personal or informal syntax or typos, for example I can meet you when be given the chance" or technical discussion, for example all of the landll methane".

5.3 Oddity

The same decision tree procedure was used to estimate a boundary between normal and substituted sentences, using the two sets of 98 sentences. This suggested a boundary value of 3.82 for the oddity measure. Using this boundary, the prediction accuracy for sentences containing substitutions is 37.8%, with a false positive rate for the normal sentences of 7%. Although the absolute predictive accuracy of the oddity measure is not high, we can compare its performance on the original sentences with the sentences in which a substitution has occurred. If the oddity measures are compared on a per-sentence basis, then 84% of the sentence pairs show an increase in the oddity measure. The measure is obviously able to detect an unusual word in a particular sentence context, but is unable to generalize this over *all* sentences.

5.4 Comparisons

None of the three measures has great accuracy by itself, so it is natural to ask whether the three measures make errors on the same sentences or on different ones. If the latter, then a combined predictor should perform much better. We build a single decision tree using the normal and substituted sentences, with the three measure values as attributes. The combined predictor has a prediction accuracy of 68% for sentences with substitutions; with a false positive rate of 16%. Practice because a message

typically consists of multiple sentences. Thresholds can be chosen to reduce the false positive rate, while detecting most of the messages containing sentences with substitutions. The difference in the performance of each of the measures suggests that part of the difficulty arises from the sheer variability of English sentences, particularly when these come from informal text where even normal grammatical irregularities are absent.

It is also clear that the boundaries derived from decision trees, using information gain as the
basic criterion, could be moved to trade off better prediction accuracy on sentences with substitutions for worse false positive rates. False positive rates may also be high because the kind of sentences used in email are much more informal and much less edited than sentences that appear in web pages.

Fig. 1. Detection performance results

Measure	Detection Rate (%)	False Positive Rate (%)	Boundary score
Sentence oddity	71	20	2.5
Left k-gram	51	28	461
Right k-gram	89	48	722
Average k-gram	51	13	418
Minimum hypernym	40	15	10
Maximum hypernym	68	31	10
Average hypernym	59	22	0

6. CONCLUSION

We have tested how word substitutions within textual communication can be detected. Our technique allows us to automatically suspicious messages, so that they can be further investigated, either by a more sophisticated data-mining techniques or manually. The task of detecting substitutions is becoming important since terrorists, criminals, spies and other adversarial parties may use substitution in order to avoidd because of the use of certain words (e.g. `bomb', `explosives', `attack', etc.). Our technique extends prior work, which was not able to detect substitutions when a word is replaced by another word.

7. REFERENCES

1. J.A. Bilmes and K. Kirchho®. Factored language models and generalized parallel backo®. In *Proceedings of HLT/NACCL*, 2003.
2. British National Corpus (BNC), 2004. www.natcorp.ox.ac.uk.
3. European Parliament Temporary Committee on the ECHELON Interception Sys-tem. Final report on the existence of a global system for the interception of private and commercial communications (ECHELON interception system), 2001.
4. SW. Fong, D.B. Skillicorn, and D. Roussinov. Detecting word substitution in adversarial communication. In *Workshop on Link Analysis, Counterterrorism and Security at SIAM International Conference on Data Mining*, submitted.
5. 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.
6. R. Ferrer i Cancho and R.V. Sol¶e. The small world of human language. *Proceedings of the Royal Society of London Series B { Biological Sciences*, pages 2261{2265, 2001.
7. H. Lee and A.Y. Ng. Spam deobfuscation using a Hidden Markov Model. In *Proceedings of the Second Conference on Email and Anti-Spam*, 2005.
8. D. Roussinov and L. Zhao. Automatic discovery of similarity relationships through web mining. *Decision Support Systems*, pages 149{166, 2003.
9. D. Roussinov, L. Zhao, and W. Fan. Mining context speci¯c similarity relationships using the World Wide Web. In *Proceedings of the 2005 Conference on Human Language Technologies*, 2005.
10. D.B. Skillicorn. Beyond keyword ¯ltering for message and conversation detec-tion. In *IEEE International*