Detecting Word Substitution in Adversarial Communication

SW. Fong and D.B. Skillicorn
School of Computing
Queen's University
{fong,skill}@cs.queensu.ca

D. Roussinov
W.P. Carey School of Business
Arizona State University
dmitri.roussinov@asu.edu

Abstract

When messages may be intercepted because they contain certain words, terrorists and criminals may replace such words by other words or locutions. If the replacement words have different frequencies from the original words, techniques to detect the substitution are known. In this paper, we consider ways to detect replacements that have *similar* frequencies to the original words. We consider the frequencies of generalized n-grams that we call k-grams, the change in frequency that results from removing the word under consideration from its context, and the change in frequency that results from replacing a word by its hypernym. In our preliminary results, none of these measures are strong individually, but together they become effective.

1 Motivation

Terrorists are aware that their communications are likely to be intercepted by systems such as Echelon [3], and would like to conceal the content of these communications as much as possible. Criminals also face similar issues since their communications may be intercepted by law enforcement.

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.

Another strategy to conceal the content of 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 [5].

However, replacement of words by words of

similar frequency becomes possible given access to a word-frequency table (for example, www.wordcount.org/main.php) or a codebook[1]. In this paper, we examine whether such substitutions are detectable by software. 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.

The contribution of this paper is the design of three measures to detect the replacement of a word by a word of similar frequency, and preliminary results about their effectiveness on a dataset of sentences taken from the Enron email dataset.

2 Related Work

The problem of detecting a word that is somehow out of context occurs in a number of settings. For example, speech recognition algorithms model the expected next word, and back up to a different interpretation when the next word becomes sufficiently unlikely [1]. This problem differs from the problem addressed here because of the strong left context that is used to decide on how unlikely the next word is, and the limited amount of resources that can be applied to detection because of the near-realtime performance requirement.

Detecting words out of context can also be used to detect (and correct) misspellings [4]. 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.

[1] although this may be difficult in practice in real time communication and in times of stress

Detecting words out of context has also been applied to the problem of spam detection. 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.

Skillicorn [8] showed that replacing words that might be on a watchlist by words with significantly different natural frequency becomes detectable, especially when the same substitution occurs in multiple texts. This is because intercepted texts, such as emails or phone calls, are actually the conglomeration of a large number of conversations. Conversations using rare words are rare; a conversation about a common topic that is expressed using a rare word therefore looks unusual. A conversation becomes more unusual either by using a substituted word with a greatly *different* frequency from the word it replaces, or because the same substitution appears in *many* messages.

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 the specific contexts was considered in [6, 7].

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 fit 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 neighbours 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 hypernym[2] 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. (This was surprising to us, and we had expected the opposite to be true; the intuition seems to be that ordinary meaningful sentences are already frequent, and hypernym replacement makes them less ordinary; whereas sentences with a substitution are already infrequent, and hypernym replacement makes them more ordinary.)

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) [2] to discard any string that appears not to be an English word, and also as a canonical source of frequencies of English words.

We extracted all strings ending with periods as possible sentences, except when the BNC corpus indicated the possibility of periods as integral parts of words, e.g. 'Mr.'. Sentences with fewer than 5 words or more than 15 words were discarded, leaving a total of 712,662 candidate sentences. A random sample of 200 sentences were drawn from this set. The size of this sample is constrained by the time taken to process the set.

Sentences containing substitutions were constructed from this set by finding the first noun that did not appear in a stopword list, and replacing it by the next most frequent noun from the BNC corpus. The stopword list in Wordnet 2.0 [3] was used.

Only sentences for which a hypernym exists in Wordnet 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.

[2]The hypernym of a noun is the noun describing a wider category of objects that include the original objects.

[3]Available from wordnet.princeton.edu

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 example, "chase the dog" occurs 9,580 times whereas "chase dog" occurs 709 times, so quoted string searches clearly do not ignore stopwords. On the other hand, the bag of words search {chase the dog} occurs 6,510,000 times while {chase dog} occurs only 6,490,000 times, which seems counterintuitive. Fourth, the order of words is significant, even in bag of word searches.

We use only the number of pages returned by Google as a surrogate for word frequency, which fails to take into account intraword frequencies within each individual document. Frequencies returned by Google should be adjusted to reflect 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.

In general, searches assume a bag of words model, that is the words of the sentence are sent to Google as individual words. When an exact search is used (a quoted search string), this will be specified.

The use of Google is only a convenience; any other source of word and sentence fragment frequencies would serve equally well. Indeed, results might be better when the source of frequencies is based more closely on the domain of discourse in the intercepted communications.

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-stopword to its left, and the first non-stopword 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 Oddity measures

When a substitution has occurred, the frequency of the entire remainder of the sentence, without the substituted word, might be expected to be high, since it is a part of an ordinary sentence that appeared in the email dataset. The frequency of the sentence containing the substituted word might be expected to be much lower, since the substituted word is unusual in the context of the remainder of the sentence.

Let f_{wo} be the frequency of the bag of words with the word under consideration omitted, and

f the frequency of the bag of words from the entire sentence. The *oddity* of a sentence can be defined as:

$$\text{oddity} \;=\; \frac{f_{wo}}{f}$$

The larger this quantity, the more likely it is that the substituted word is unusual.

The problem with the oddity measure is that, when the substituted word is common, the numerator and denominator frequencies become more and more similar, and so it becomes harder and harder to detect the presence of an unusual word. We have experimented with a variety of normalization terms, for example multiplying the oddity by the frequency of the omitted word, but these do not seem to improve results.

3.5 Hypernym measures

The hypernym of a word is another word that describes a more-general class of objects of which the initial word is an example. For example, 'vehicle' is a hypernym for 'car', 'train', and 'sleigh'. When a substitution has occurred, the hypernym of the substituted word should seem more appropriate (and conversely when a word is already appropriate, its hypernym may not be as appropriate).

Let f be the frequency of the bag of words containing the substituted word, and f_H be the frequency of the bag of words with the substituted word replaced by its hypernym (obtained via Wordnet). Then we compute a score

$$\text{hypernym oddity} \;=\; f_H - f$$

which takes into account the amount by which the sentence with substitution and sentence with hypernym frequencies differ. We expect this measure to be close to zero or negative for normal sentences, but positive for sentences that contain unusual words. (We could have defined this measure as a ratio, but the boundary between classes occurs when the frequencies are of similar magnitude, so the interesting cases would be very close to 1 and so hard to work with.)

There are two obvious problems with using hypernyms. First, as the hierarchy of hypernyms is climbed, the hypernyms are increasingly likely to become technical terms which are not actually used in sentences. For example, the hypernym of 'car' is 'motor vehicle'; and its hypernym is 'self-propelled vehicle'. The base frequencies of these phrases are: 'car': 395M; 'motor vehicle': 18.7M, and 'self-propelled vehicle": 72,300. The reduced frequencies of the individual locutions will drag down the frequencies of the bags of words containing them.

Second, many words have several hypernyms depending on the sense in which they are being used. Without semantic information, it is not straightforward to choose the appropriate hypernym.

3.6 Examples

To illustrate the issues that may arise, we consider two examples, one a sentence where the substitution is easy to detect, and the other a sentence where the substitution is difficult to detect.

Consider this sentence: "copyright 2001 southwest airlines co all rights reserved". The first noun is 'copyright' and this is replaced by the noun 'toast' which has almost the same frequency in the BNC corpus. So the sentence we consider is "toast 2001 southwest airlines co all rights reserved", which any human would immediately detect as unusual.

The k-gram around the substituted word is "toast 2001" and the frequency of this exact string is 0. This is a commonly observed pattern for k-grams containing substitutions; rather than occurring only infrequently, they tend not to occur at all. (The equivalent k-gram from the original sentence, "copyright 2001", on the other hand, has frequency 3,120,000.)

The oddity of the sentence with the substitution is the ratio of the frequency of the bag of words {2001 southwest airlines co all rights reserved} (33,400) to the frequency of the bag of words {toast 2001 southwest airlines co all rights

reserved} (1620), giving an oddity of 20.62. (The corresponding oddity of the original sentence is 33,400/16,100 = 2.07.)

The automatically chosen hypernym for 'toast' is 'bread'. The frequency of the bag of words {bread 2001 southwest airlines co all rights reserved} is 4120 , while the frequency of the bag of words {toast 2001 southwest airlines co all rights reserved} is 1620. The hypernym oddity is therefore 2500. (The corresponding hypernym oddity of the original sentence is -6840).

For this sentence, all of the measures give strong indications of the presence of an unusual substitution.

Now we consider a more difficult sentence. The original sentence is "please try to maintain the same seat each class". The noun 'seat' is replaced by 'play', the next most frequent noun in the BNC corpus. So the sentence we consider is "please try to maintain the same play each class" which a human would certainly consider unusual, but which might perhaps make sense in certain settings, for example a drama school.

The k-gram around the substituted word is 'maintain the same play each class" and the frequency of this exact string is 0. (The equivalent k-gram from the original sentence, "maintain the same seat each class" also has frequency 0.)

The oddity of the sentence with the substitution is the ratio of the frequency of the bag of words {please try to maintain the same each class} (749,000) to the frequency of the bag of words {please try to maintain the same play each class} (528,000), giving an oddity of 1.42. (The corresponding oddity of the original sentence is 749,000/125,000 = 5.99.) The problem here is that 'play' is a common word (and has several different senses) so adding it to the search terms results in only a small decrease in frequency.

The automatically chosen hypernym for 'play' is 'dramatic composition', a typical illustration of a hypernym that is quite technical, and therefore not in common use. The frequency of the bag of words {please try to maintain the same play each class} is 528,000 , while the frequency of the bag of words {please try to maintain the same dramatic composition each class} is 32,400. The hypernym oddity is therefore 274,000. (The corresponding hypernym oddity of the original sentence is -495,600).

All three of these measures provide hints about the presence of an unusual substitution. However, especially with the knowledge of the corresponding measures for the original, normal sentence, the results are not compelling. Although the k-gram measure correctly indicates that the k-gram never occurs, nor does the equivalent k-gram for the original sentence. The oddity for the sentence with the substitution is low (and much lower than the oddity of the original sentence).

4 Results and Discussion

4.1 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.

4.2 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 landfill methane".

4.3 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.

4.4 Hypernym oddity

Once again a decision tree process was used to estimate a boundary between normal and substituted sentences. This suggested a boundary value of -3. This accords well with intuition, which would have suggested a boundary of 0, since we expect sentences containing a substitution to have increased frequency in their hypernym versions.

Using this boundary, the prediction accuracy for sentences containing substitutions is 56%, with a false positive rate for normal sentences of 22%.

Once again, the absolute predictive accuracy of the hypernym oddity measure is not high. Again, the comparisons on a per-sentence basis show that the hypernym oddity increases for 76% of the sentence pairs. Since choosing the 'wrong' hypernym explains much of the poor performance of this measure, we are exploring ways to compensate, for example by trying all possible hypernyms.

4.5 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%.

Such a prediction accuracy can be useful in 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.

The results are shown in the following table:

Technique	Prediction Accuracy %	False Positive %
k-grams	81	47
Oddity	38	7
Hypernym oddity	56	22
Combined	68	16

5 Conclusions

We have presented preliminary results for some measures to detect word substitutions in sentences. We assume such substitutions replace words with words of similar frequency, so that 1-gram frequency counting techniques cannot detect the replacements. Our measures are designed to provide information about semantic discontinuities caused by substitutions indirectly, via string frequencies obtained from Google. Of the three measures considered, k-gram frequency seems to perform best, but it is limited by the fact that many k-grams never occur, even in normal text. Each of the measures performs poorly on its own, although we are exploring improvements to each, but together they begin to become practically effective.

References

[1] J.A. Bilmes and K. Kirchhoff. Factored language models and generalized parallel backoff. 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 System. Final report on the existence of a global system for the interception of private and commercial communications (echelon interception system), 2001.

[4] 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.

[5] M.L. Newman, J.W. Pennebaker, D.S. Berry, and J.M. Richards. Lying words: Predicting deception from linguistic style. *PSPB*, 29:665–675, 2003.

[6] D. Roussinov, W. Fan, and L. Zhao. Discrimination based learning approach to information retrieval. In *Proceedings of the 2005 Conference on Human Language Technologies*, 2005.

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

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