Using A Probabilistic Model Of Context To Detect Word Obfuscation

Sanaz Jabbari, Ben Allison, Louise Guthrie

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello, Sheffield, S1 4DP
{s.jabbari, b.allison, L.guthrie}@dcs.shef.ac.uk

Abstract

This paper proposes a distributional model of word use and word meaning which is derived purely from a body of text, and then applies this model to determine whether certain words are used in or out of context. We suggest that we can view the contexts of words as multinomially distributed random variables. We illustrate how using this basic idea, we can formulate the problem of detecting whether or not a word is used in context as a likelihood ratio test. We also define a measure of semantic relatedness between a word and its context using the same model. We assume that words that typically appear together are related, and thus have similar probability distributions and that words used in an unusual way will have probability distributions which are dissimilar from those of their surrounding context. The relatedness of a word to its context is based on Kullback-Leibler divergence between probability distributions assigned to the constituent words in the given sentence. We employed our methods on a defense-oriented application where certain words are substituted with other words in an intercepted communication.

1. Introduction

The work presented in this paper defines a probabilistic model of context and demonstrates how that model might be applied to the problem of detecting *textual obfuscations*. By textual obfuscation we mean the substitution of one word instead of another in a sentence. Most automatic identification of potentially criminal or dangerous communication are currently based on simple keyword–spotting, capable of recognising and suspending messages that contain certain predetermined, "red flagged" words such as *bomb* and *poison*. However, an obvious bypass mechanism for the sender (or speaker) is to avoid using red–flagged words by replacing them with apparently innocent words (for example, *heroin* may be replaced with *horse*).

A simple keyword-based system would fail to spot such substitutions. This paper proposes a probabilistic model of context and suggests two methods that might be used to detect such substitutions. One initial experiment for this task is the following: given a sentence, and some word within that sentence marked for consideration, does the word fit the context of that sentence? For example, we might be asked with determining whether the word *dancers* is a substitution in the following sentences:

- Perhaps no ballet has ever made the same impact on *dancers* and audience as Stravinsky's "Rite of Spring".
- He remembered sitting on the wall with a cousin, watching the German *dancers* fly over.

The task is to determine that sentence one is a plausible context for *dancers*, whereas sentence two is not.

The rest of the paper is presented as follows: section 2. gives a brief overview of work in the language processing literature which attempts similar problems. Section 3. describes our model of context, and section 4. presents two methods that use our model to judge if a word is used out of its context. Section 5. describes our experimental setup and our results. Finally section 6. ends with some concluding remarks and our future direction.

2. Background

(Fong and Skillcorn, 2006) address a very similar problem of detecting word substitution in intercepted communication. One of the major challenges in this problem is using appropriate data for evaluation. Since the problem is envisaged as a defense application, getting hold of the real data is extremely difficult; on the other hand, manual creation of the data is sensitive to the subjectivity of the substitutions. (Fong and Skillcorn, 2006) use the Enron email dataset as their test, and gather sentences in the messages, replacing the first noun of each sentence with the noun that has the closest frequency list of the nouns in the British National Corpus. They use three measures for detecting the substitutions; an oddity measure compared to data found on th Web, a semantic measure using WordNet (Fellbaum, 1998), and finally a measure that counts the frequency of the left and right bigrams around the target word.

Word obfuscation detection is one of the many natural langugae processing tasks that can benefit from characterising the contexts a word or a phrase are typically used in. Using a plausible model of context, an intelligent spelling (and grammar) checker could select correctly spelled but incorrect words by the degree to which they fit in their surrounding context. A summariser could keep its context as close as possible to the one induced by the original text. In language generation for dialog systems, in order to create a natural dialogue, a paraphrase generator with a good model of context can be used to ensure that the new context is as similar as possible, if not equivalent to the original one. Even more ambitiously, for evaluating the output of a machine translation system, given that the same piece of text in both languages should share the same context, the degree of overlap between the two contexts could be used as one of the measures of translation quality. Specifically, a model of context can be used for translating a polysemous word to

its correct word in the target language. Furthermore, since a polysemous word can be disambiguated by its surrounding context, models of context can be used in all tasks that are designed to evaluate the word sense disambiguation systems. For instance for the Lexical Substitution task which was introduced in Semeval 2007 (McCarthy and Navigli, 2007), systems were asked to find a set of words that could be substituted with a target word in its given context (Giuliano et al., 2007; Zhao et al., 2007; Yuret, 2007).

The approaches to this problem could be divided to two groups. The first method assumes that if a word is used in context, it is formally and logically consistent with its context. For instance the Generative Lexicon (Pustejovsky, 1991) defines a coherent context by propagating type constraints and type coercions through entries in a lexicon, and thus allowed words are those which satisfy the constraints defined by the surrounding context. A similar approach is that of lexical chains: if a word can be linked to the chain defined by its context, then it contributes to the coherence of the text, and hence can be regarded as a consistent element of that context. In this vein, (Hirst and St-Onge, 1997) demonstrates the construction of chains from WordNet defines a coherence relation with respect to WordNet (Fellbaum, 1998) to detect spelling errors.

While such approaches arguably provide much better insight into the linguistic processing capabilities of humans, as a practical approach they are handicapped by relying on large amounts of carefully hand crafted resources.

Another group of approaches uses the concept of semantic relatedness. In this paradigm, words and their contexts are treated as two separate entities defined in a common space. Judging their consistency can then be accomplished by measuring their distance in this space. These measures of distance rely either on hand–crafted semantic networks (e.g. Wordnet, Roget's thesaurus , etc.) (Slator, 1989; Leacock et al., 1998; Resnik, 1999; Jarmasz and Szpakowicz, 2003) or on information found in corpora (Lee, 1997; Lee and Pereira, 1999; Lee, 1999; Blagoev. and Mulloni, 2007). Despite the accuracy and accessibility of information in hand–crafted resources, there are many concerns with the coverage of the resources, and any distance measure defined in terms of such a resource requires further heuristics to be defined over the resource.

Corpus–based approaches, on the other hand, assume that a word's meaning can be captured in its use. These methods employ a "word–space model" into which a word and its context are mapped, and their proximity within this space represents their semantic similarity. Such spaces are often constructed from term-by-term (co-occurrence) matrices, and the set of words selected as the features of this space and what sort of information is countable for a given dimension are the major distinguishing factors in such models (Sahlgren, 2006). For instance, some models consider unconditional co-occurrence of words, while other models consider co-occurrences within certain grammatical relationships (Padó and Lapata, 2007; Rothenhusler. and Schütze, 2007). Other models consider dimension–reducing transformations of the original co-occurrence matrix, such as factor analysis or Singular Value Decomposition (Landauer T. K. and D., 1998).

Even assuming that the meanings of individual words are successfully represented within some word space, representing their context in the same space more of a challenge. For example, (Schütze, 1998) defines "context vectors" to be the centroid of the vectors of content words in the context. (Gliozzo, 2005) represents both words and contexts in a domain space, and uses domain vectors to represent words and contexts.

3. A Probabilistic model of context

In this work, we choose to model contexts probabilistically. Any sentence (or any other unit of context, e.g. two nouns either side of w) containing w could be regarded as a random sample from some process depending on a distribution p_w, and these different processes allow us to capture different aspects of the context of w.

We propose the following simple model: let the random variable C be multinomial, such that $\{C_1...C_v\}$ represent counts of words $\{w_1...w_v\}$ in some context (this could embrace either the whole vocabulary or some discriminatory subset). Consider that C represents possible outcomes of the experiment: select a word at random according to the distribution $p_w(w_j)$, and do so n times for a context of length n. The distribution over C is then multinomial with parameters (θ, n), where $\theta = \{\theta_1...\theta_v\}$ is simply $\theta_j = p_w(w_j)$ for each w_j.

Proper estimation for the distributions p_w is a difficult matter, however a simple estimate can be derived as follows. Let $p_{w_i}(w_j)$be the probability that a randomly selected word from w_i's context is w_j. Then let n_{ij} be the count of the co-occurrence (w_i, w_j) in the background corpus and $n_{\bullet\bullet}$be the total count of all co-occurrences which contain w_i. A possible estimate is then:

$$p_{w_i}(w_j) = \frac{n_{ij} + 1}{n_{\bullet\bullet} + |V|} \quad (1)$$

which is equivalent to assuming that the probabilities over the w_js for some fixed w_i arise from a Dirichlet prior with each $\alpha_j = 1$, or alternatively that each co-occurrence occurs once more than the actual count in the corpus. In this work, we derive the relevant counts from the English Gigaword, a 1.5 billion word corpus of newswire text.

Given estimated distributions of the form above, we employed three approaches to determine whether a given word in a sentence is fake/substituted or not:

1. Testing the hypothesis that the context in question, c, was sampled from a distribution with parameters $\theta_j = p_{w_i}(w_j)$
2. Quantifying the semantic similarity between the word w_i and other words in its its context

We elaborate on these methods in the section below.

4. Identifying improper contexts based on our model

After associating each word a corresponding distribution as described in section 3., we propose three methods for deciding whether a word is an obfuscation.

4.1. Method 1: A likelihood ratio hypothesis test

After assigning a probability distribution to target word (e.g. *dancers*), we would like to measure the probability that the candidate context is generated from this distribution. Each word is associated with a multinomial distribution with parameters $\{\theta_1 ... \theta_v\}$, which, as described above, are simply $\theta_j = p_w(w_j)$ for each possible outcome w_j. A word's context is represented by the multivariate random variable $C = \{C_1 ... C_v\}$, which are counts of the number of times word w_j occurred in the context for each w_j. We then wish to test the hypothesis that the particular context, c, is a sample from a multinomial distribution with parameters $\theta = \{\theta_1 ... \theta_v\}$, where each $\theta_j = p_w(w_j)$.

The probability of c being sampled from a distribution with these parameters can be written as follows:

$$p(c;\, \theta) \propto \prod_i \theta_i^{c_i} \tag{2}$$

Where we omit the multinomial coefficient since it does not depend upon θ.

However, if we were to compare this quantity across contexts, we would observe that longer contexts had a much lower probability for any θ; thus we also compute the probability of c given some alternate θ', and compare the ratio of these quantities across contexts: this allows a roughly length–independent comparison of the quantity of interest between contexts. This is analogous to a likelihood ratio hypothesis test, where the likelihood of some observations is compared under two different hypotheses.

The alternate parameters θ' are calculated as follows: θ_j corresponds to $p_{w_i}(w_j)$ for some w_i, which is in turn estimated based upon counts of the co-occurrence of the pair (w_i, w_j) (or the pair (w_j, w_i)) which we denoted n_{ij}. If we sum the quantity n_{ij} over all i, and use this to estimate the distribution $p_{w_\bullet}(w_j)$ this distribution corresponds in a sense to the "global" distribution of contexts (or more formally, the maximum likelihood estimate for a distribution from which we hypothesise that *all* contexts are sampled). The estimates for the θ'_j are thus:

$$\theta'_j = \frac{n_{\bullet j}}{n_{\bullet\bullet}} \tag{3}$$

Where the bullet $\bullet$ once again denotes "all".

For example, the probability that the sentence "He remembered sitting on the ...over" was sampled from a distribution with the same parameters as those of the typical contexts of *dancers*, is calculated as:

$$p_{dancers}(remebered) \times ... \times p_{dancers}(fly) \tag{4}$$

This is compared with the probability that c was sampled from a distribution $p_\bullet(c)$ with parameters θ', to yield the ratio:

$$\frac{p_{dancers}(remembered) \times ... \times p_{dancers}(fly)}{p_\bullet(remembered) \times ... \times p_\bullet(fly)} \tag{5}$$

To make a decision, we judge that all contexts for which this ratio is suitably small are substitutions.

4.2. Method 2: A semantic similarity measure for words

The assumption behind this method is that a substituted word is not semantically related to its candidate context. We propose to use a measure of semantic similarity based on the distributions $p_{w_i}(w_j)$ as estimated above, and envisage that this has applications beyond the present problem. However, for the purposes of this paper we show how this measure has application to the problem at hand.

Each word w_i is associated with a distribution p_{w_i} over other words; the distribution represents the probability that a randomly selected word from the context of w_i will be w_j. Given two words w_1 and w_2, we could look to quantify the relatedness between them as a function of the "similarity" between the distributions p_{w_1} and p_{w_2}: the standard measure for this is the Kullback-Leibler divergence.

If we denote by X the random variable taking values which are possible co-occurring words (w_j from above), then the divergence between p_{w_1} and p_{w_2}, denoted $D(p_{w_1}||p_{w_2})$, is:

$$D(p_{w_1}||p_{w_2}) = \sum_{x \in X} p_{w_1}(x) log \frac{p_{w_1}(x)}{p_{w_2}(x)} \tag{6}$$

The relatedness of the target word w (e.g. *dancers*) to its context can then be judged by the average divergence between p_w and the distributions $p_{w_1} ... p_{w_n}$ for the n words in the context of w:

$$\text{Relatedness}(w,\, \text{context}(w)) = \frac{\sum_i D(w, w_i)}{n}$$

5. Experiments and Results

We use 1.4 billion words of English Gigaword v.1, a newswire corpus collected from four different sources, as data to derive counts for the estimates. To ensure no overlap between training and testing, test sentences are built from the BNC. Four test-sets were created, where for each a word (e.g. *dancer*) was chosen and 500 sentences containing the selected word were randomly selected. These 500 sentences are considered to be the normal use of the word. We then choose another word (e.g. *bomber*), unrelated semantically to the first word (but with the same part-of-speech tag) from the vocabulary and 500 of its sentences were taken randomly from the BNC. The second word was replaced with the first word. The result is another 500 sentences, containing the first word, but in an abnormal context. The list of the substitutions is listed in table 1.

In this work, stopwords are discarded from both the test and the training sets. All probability distributions are defined over 2000 most frequent non-stop words in the Gigaword. The results are shown in table 2.

6. Discussion

The results show that as the size of the background data for a given word increases, the performance of both methods generally improves (always in case of our first method); this is not a surprising effect, since the more data should should allow more accurate parameter estimation. However, the likelihood–ratio method requires estimates of two distributions; the first is global, and thus by definition parameters are as accurate as possible given the whole training data.

Old Word	Target Word	Target word frequency
bombers	dancers	4714
president	chicken	14025
gun	cup	237080
war	championship	778595

Table 1: List of the substitutions which comprise the test data

Word	Likelihood Ratio	Semantic Similarity
dancers	0.778	0.787
chicken	0.869	0.84
cup	0.93	0.756
championship	0.94	0.867

Table 2: Results of the two methods. Scores are F-Measure for the anomalous category

The second set of parameters depends only upon occurrences of the target word, and thus more instances of the target word directly affects parameter estimation and thus performance.

The similarity method, on the other hand, requires parameter estimation for many words besides the target word (all other words in the context); thus performance also depends on the quality of these estimates, which do not improve with more instances of the target word. In both cases, despite the relatively simple model, the results above show that some notion of context is being captured by the distributions, and that that notion is useful in detecting obfuscated words.

For our future work, we will create a more principled way of creating our data set, by replacing words from one semantic category with another with respect to a lexical resource such as Longman Dictionary of Contemporary English (Procter, 1978). We would like to test probablistic models that capture more of our intuitions of the language and compare them with the simple multinomial model presented in this work.

7. References

V. Pekar. R. Mitkov. D. Blagoev. and A. Mulloni. 2007. Finding translations for low-frequency words in comparable corpora. In *In Sixth International and Interdisciplinary Conference on Modeling and Using Context*.

Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press, May.

S. W. Fong and D. B. Skillcorn. 2006. Detecting word substitution in adversarial communication.

Claudio Giuliano, Alfio Gliozzo, and Carlo Strapparava. 2007. Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, June.

Alfio Massimilliano Gliozzo. 2005. *Semantic Domains in Computational Linguistics*. Ph.D. thesis.

G. Hirst and D. St-Onge. 1997. Lexical chains as representation of context for the detection and correction malapropisms.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's thesaurus and semantic similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP-03)*, pages 212–219.

Foltz P. W. Landauer T. K. and Laham D. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Claudia Leacock, George A. Miller, and Martin Chodorow. 1998. Using corpus statistics and wordnet relations for sense identification. *Comput. Linguist.*, 24(1):147–165.

Lillian Lee and Fernando Pereira. 1999. Distributional similarity models: clustering vs. nearest neighbors. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 33–40. Association for Computational Linguistics.

Lillian Jane Lee. 1997. *Similarity-based approaches to natural language processing*. Ph.D. thesis, Cambridge, MA, USA.

Lillian Lee. 1999. Measures of distributional similarity. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 25–32.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, June.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Comput. Linguist.*, 33(2):161–199.

Paul Procter. 1978. *Longman's Dictionary of Contemporary English*. Longman Group Limited.

James Pustejovsky. 1991. The generative lexicon. *Comput. Linguist.*, 17(4):409–441.

Philip Resnik. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130.

K. Rothenhusler. and H. Schütze. 2007. Part of speech filtered word spaces. In *In Sixth International and Interdisciplinary Conference on Modeling and Using Context*.

Magnus Sahlgren. 2006. *The Word-Space Model: using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Comput. Linguist.*, 24(1):97–123.

B. M. Slator. 1989. Extracting lexical knowledge from dictionary text. *SIGART Bull.*, (108):173–174.

Deniz Yuret. 2007. Ku: Word sense disambiguation by substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June. Association for Computational Linguistics.

Shiqi Zhao, Lin Zhao, Yu Zhang, Ting Liu, and Sheng Li. 2007. Hit: Web based scoring method for english lexical substitution. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics, June.