Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 199-206.

Entropy Rate Constancy in Text

Dmitriy Genzel and **Eugene Charniak**
Brown Laboratory for Linguistic Information Processing
Department of Computer Science
Brown University
Providence, RI, USA, 02912
{dg,ec}@cs.brown.edu

Abstract

We present a *constancy rate principle* governing language generation. We show that this principle implies that local measures of entropy (ignoring context) should increase with the sentence number. We demonstrate that this is indeed the case by measuring entropy in three different ways. We also show that this effect has both lexical (*which* words are used) and non-lexical (*how* the words are used) causes.

1 Introduction

It is well-known from Information Theory that the most efficient way to send information through noisy channels is at a constant rate. If humans try to communicate in the most efficient way, then they must obey this principle. The communication medium we examine in this paper is text, and we present some evidence that this principle holds here.

Entropy is a measure of information first proposed by Shannon (1948). Informally, entropy of a random variable is proportional to the difficulty of correctly guessing the value of this variable (when the distribution is known). Entropy is the highest when all values are equally probable, and is lowest (equal to 0) when one of the choices has probability of 1, i.e. deterministically known in advance.

In this paper we are concerned with entropy of English as exhibited through written text, though these results can easily be extended to speech as well. The random variable we deal with is therefore a unit of text (a word, for our purposes[1]) that a random person who has produced all the previous words in the text stream is likely to produce next. We have as many random variables as we have words in a text. The distributions of these variables are obviously different and depend on all previous words produced. We claim, however, that the entropy of these random variables is on average the same[2].

2 Related Work

There has been work in the speech community inspired by this *constancy rate principle*. In speech, distortion of the audio signal is an extra source of uncertainty, and this principle can by applied in the following way:

A given word in one speech context might be common, while in another context it might be rare. To keep the entropy rate constant over time, it would be necessary to take more time (i.e., pronounce more carefully) in less common situations. Aylett (1999) shows that this is indeed the case.

It has also been suggested that the principle of constant entropy rate agrees with biological evidence of how human language processing has evolved (Plotkin and Nowak, 2000).

Kontoyiannis (1996) also reports results on 5 consecutive blocks of characters from the works

[1]It may seem like an arbitrary choice, but a word is a natural unit of length, after all when one is asked to give the length of an essay one typically chooses the number of words as a measure.

[2]Strictly speaking, we want the cross-entropy between all words in the sentences number n and the true model of English to be the same for all n.

of Jane Austen which are in agreement with our principle and, in particular, with its corollary as derived in the following section.

3 Problem Formulation

Let $\{X_i\}, i = 1 \ldots n$ be a sequence of random variables, with X_i corresponding to word w_i in the corpus. Let us consider i to be fixed. The random variable we are interested in is Y_i, a random variable that has the same distribution as $X_i|X_1 = w_1, \ldots, X_{i-1} = w_{i-1}$ for some fixed words $w_1 \ldots w_{i-1}$. For each word w_i there will be some word w_j, $(j \leq i)$ which is the starting word of the sentence w_i belongs to. We will combine random variables $X_1 \ldots X_{i-1}$ into two sets. The first, which we call C_i (for context), contains X_1 through X_{j-1}, i.e. all the words from the preceding sentences. The remaining set, which we call L_i (for local), will contain words X_j through X_{i-1} . Both L_i and C_i could be empty sets. We can now write our variable Y_i as $X_i|C_i, L_i$.

Our claim is that the entropy of Y_i , $H(Y_i)$ stays constant for all i. By the definition of relative mutual information between X_i and C_i,

$$\begin{aligned} H(Y_i) &= H(X_i|C_i, L_i) \\ &= H(X_i|L_i) - I(X_i|C_i, L_i) \end{aligned}$$

where the last term is the mutual information between the word and context given the sentence. As i increases, so does the set C_i. L_i, on the other hand, increases until we reach the end of the sentence, and then becomes small again. Intuitively, we expect the mutual information at, say, word k of each sentence (where L_i has the same size for all i) to increase as the sentence number is increasing. By our hypothesis we then expect $H(X_i|L_i)$ to increase with the sentence number as well.

Current techniques are not very good at estimating $H(Y_i)$, because we do not have a very good model of context, since this model must be mostly semantic in nature. We have shown, however, that if we can instead estimate $H(X_i|L_i)$ and show that it increases with the sentence number, we will provide evidence to support the constancy rate principle.

The latter expression is much easier to estimate, because it involves only words from the beginning of the sentence whose relationship is largely local and can be successfully captured through something as simple as an n-gram model.

We are only interested in the mean value of the $H(X_j|L_j)$ for $w_j \in S_i$, where S_i is the ith sentence. This number is equal to $\frac{1}{|S_i|}H(S_i)$, which reduces the problem to the one of estimating the entropy of a sentence.

We use three different ways to estimate the entropy:

- Estimate $H(S_i)$ using an n-gram probabilistic model
- Estimate $H(S_i)$ using a probabilistic model induced by a statistical parser
- Estimate $H(X_i)$ directly, using a non-parametric estimator. We estimate the entropy for the beginning of each sentence. This approach estimates $H(X_i)$, not $H(X_i|L_i)$, i.e. ignores not only the context, but also the local syntactic information.

4 Results

4.1 N-gram

N-gram models make the simplifying assumption that the current word depends on a constant number of the preceding words (we use three). The probability model for sentence S thus looks as follows:

$$\begin{aligned} P(S) &= P(w_1)P(w_2|w_1)P(w_3|w_2w_1) \\ &\times \prod_{i=4}^{n} P(w_n|w_{n-1}w_{n-2}w_{n-3}) \end{aligned}$$

To estimate the entropy of the sentence S, we compute $\log P(S)$. This is in fact an estimate of cross entropy between our model and true distribution. Thus we are overestimating the entropy, but if we assume that the overestimation error is more or less uniform, we should still see our estimate increase as the sentence number increases.

Penn Treebank corpus (Marcus et al., 1993) sections 0-20 were used for training, sections 21-24 for testing. Each article was treated as a separate text, results for each sentence number were

grouped together, and the mean value reported on Figure 1 (dashed line). Since most articles are short, there are fewer sentences available for larger sentence numbers, thus results for large sentence numbers are less reliable.

The trend is fairly obvious, especially for small sentence numbers: sentences (with no context used) get harder as sentence number increases, i.e. the probability of the sentence given the model decreases.

4.2 Parser Model

We also computed the log-likelihood of the sentence using a statistical parser described in Charniak (2001)[3]. The probability model for sentence S with parse tree T is (roughly):

$$P(S) = \prod_{x \in T} P\left(x | parents(x)\right)$$

where $parents(x)$ are words which are parents of node x in the the tree T. This model takes into account syntactic information present in the sentence which the previous model does not. The entropy estimate is again $\log P(S)$. Overall, these estimates are lower (closer to the true entropy) in this model because the model is closer to the true probability distribution. The same corpus, training and testing sets were used. The results are reported on Figure 1 (solid line). The estimates are lower (better), but follow the same trend as the n-gram estimates.

4.3 Non-parametric Estimator

Finally we compute the entropy using the estimator described in (Kontoyiannis et al., 1998). The estimation is done as follows. Let T be our training corpus. Let $S = \{w_1 \dots w_n\}$ be the test sentence. We find the largest $k \leq n$, such that sequence of words $w_1 \dots w_k$ occurs in T. Then $\frac{\log S}{k}$ is an estimate of the entropy at the word w_1. We compute such estimates for many first sentences, second sentences, etc., and take the average.

[3]This parser does not proceed in a strictly left-to-right fashion, but this is not very important since we estimate entropy for the whole sentence, rather than individual words

For this experiment we used 3 million words of the Wall Street Journal (year 1988) as the training set and 23 million words (full year 1987) as the testing set[4]. The results are shown on Figure 2. They demonstrate the expected behavior, except for the strong abnormality on the second sentence. This abnormality is probably corpus-specific. For example, 1.5% of the second sentences in this corpus start with words "the terms were not disclosed", which makes such sentences easy to predict and decreases entropy.

4.4 Causes of Entropy Increase

We have shown that the entropy of a sentence (taken without context) tends to increase with the sentence number. We now examine the causes of this effect.

These causes may be split into two categories: lexical (which words are used) and non-lexical (how the words are used). If the effects are entirely lexical, we would expect the per-word entropy of the closed-class words not to increase with sentence number, since presumably the same set of words gets used in each sentence. For this experiment we use our n-gram estimator as described in Section 4.2. We evaluate the per-word entropy for nouns, verbs, determiners, and prepositions. The results are given in Figure 3 (solid lines). The results indicate that entropy of the closed class words increases with sentence number, which presumably means that non-lexical effects (e.g. usage) are present. We also want to check for presence of lexical effects. It has been shown by Kuhn and Mohri (1990) that lexical effects can be easily captured by caching. In its simplest form, caching involves keeping track of words occurring in the previous sentences and assigning for each word w a caching probability $P_c(w) = \frac{C(w)}{\sum_w C(w)}$, where $C(w)$ is the number of times w occurs in the previous sentences. This probability is then mixed with the regular probability (in our case - smoothed trigram) as follows:

$$P_{mixed}(w) = (1 - \lambda) P_{ngram}(w) + \lambda P_c(w)$$

[4]This is not the same training set as the one used in two previous experiments. For this experiment we needed a larger, but similar data set

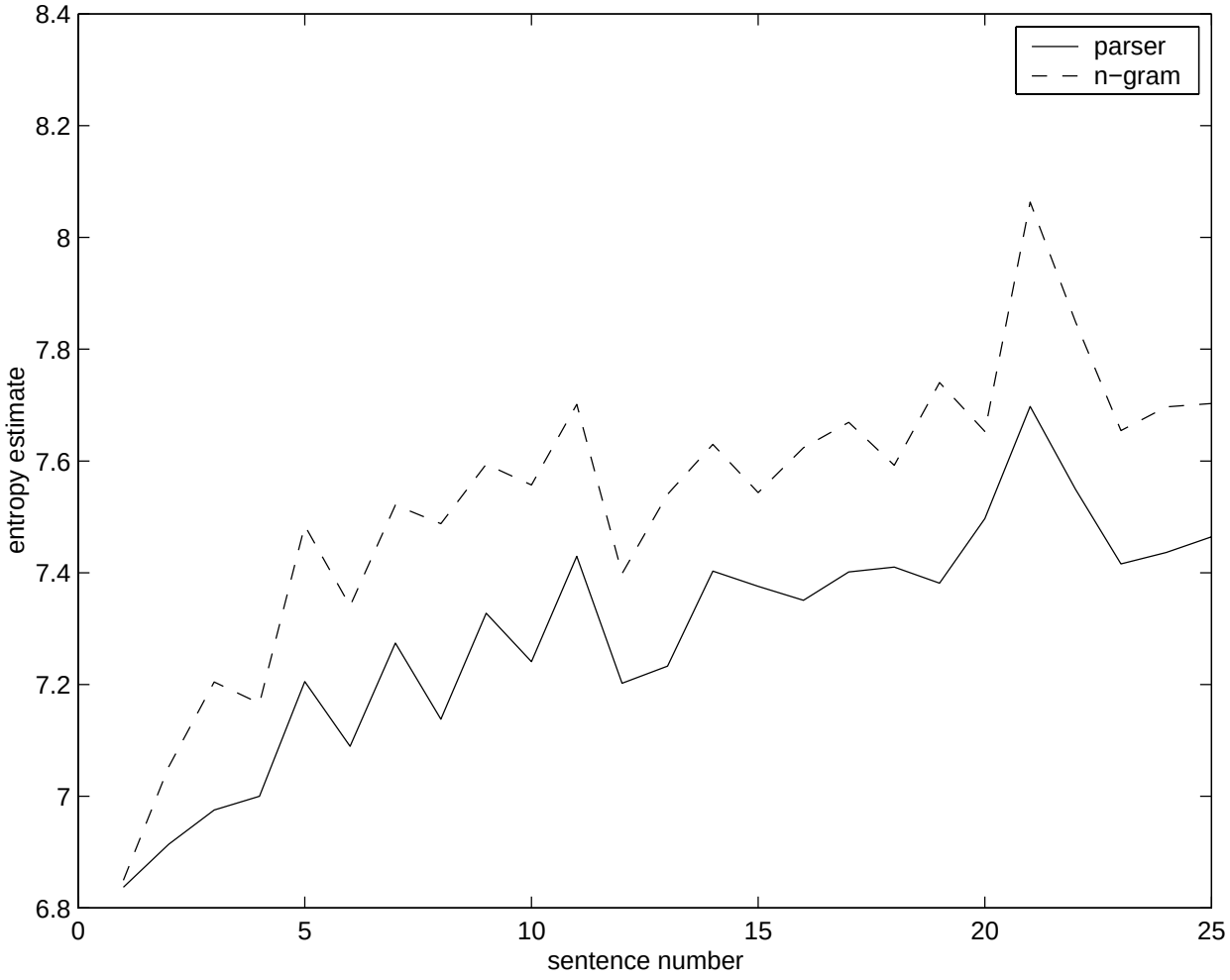

Figure 1: N-gram and parser estimates of entropy (in bits per word)

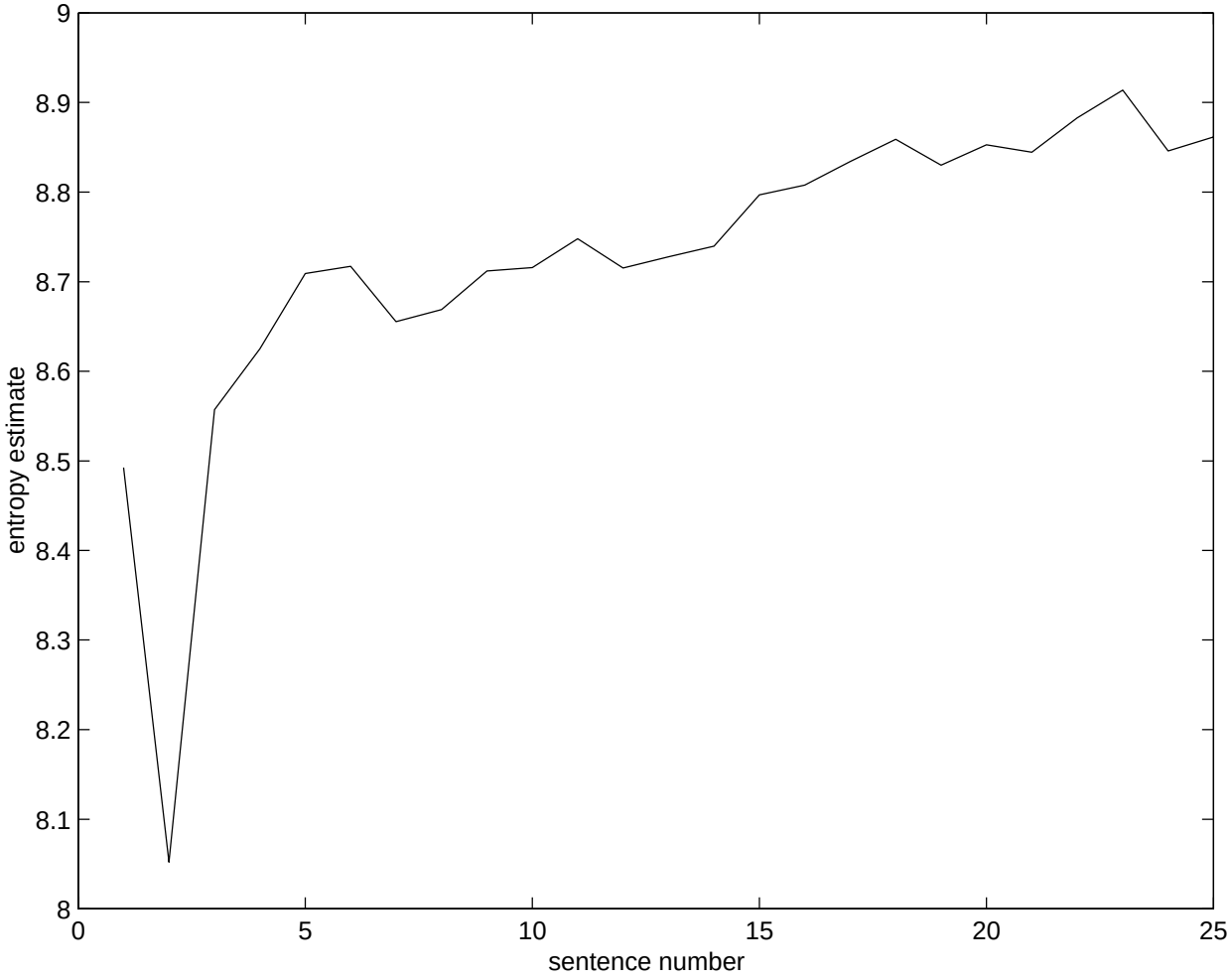

Figure 2: Non-parametric estimate of entropy

where λ was picked to be 0.1. This new probability model is known to have lower entropy. More complex caching techniques are possible (Goodman, 2001), but are not necessary for this experiment.

Thus, if lexical effects are present, we expect the model that uses caching to provide lower entropy estimates. The results are given in Figure 3 (dashed lines). We can see that caching gives a significant improvement for nouns and a small one for verbs, and gives no improvement for the closed-class parts of speech. This shows that lexical effects are present for the open-class parts of speech and (as we assumed in the previous experiment) are absent for the closed-class parts of speech. Since we have proven the presence of the non-lexical effects in the previous experiment, we can see that both lexical and non-lexical effects are present.

5 Conclusion and Future Work

We have proposed a fundamental principle of language generation, namely the *entropy rate constancy* principle. We have shown that entropy of the sentences taken without context increases with the sentence number, which is in agreement with the above principle. We have also examined the causes of this increase and shown that they are both lexical (primarily for open-class parts of speech) and non-lexical.

These results are interesting in their own right, and may have practical implications as well. In particular, they suggest that language modeling may be a fruitful way to approach issues of contextual influence in text.

Of course, to some degree language-modeling caching work has always recognized this, but this is rather a crude use of context and does not address the issues which one normally thinks of when talking about context. We have seen, however, that entropy measurements can pick up much more subtle influences, as evidenced by the results for determiners and prepositions where we see no caching influence at all, but nevertheless observe increasing entropy as a function of sentence number. This suggests that such measurements may be able to pick up more obviously semantic contextual influences than simply the repeating words captured by caching models. For example, sentences will differ in how much useful contextual information they carry. Are there useful generalizations to be made? E.g., might the previous sentence always be the most useful, or, perhaps, for newspaper articles, the first sentence? Can these measurements detect such already established contextual relations as the given-new distinction? What about other pragmatic relations? All of these deserve further study.

6 Acknowledgments

We would like to acknowledge the members of the Brown Laboratory for Linguistic Information Processing and particularly Mark Johnson for many useful discussions. Also thanks to Daniel Jurafsky who early on suggested the interpretation of our data that we present here. This research has been supported in part by NSF grants IIS 0085940, IIS 0112435, and DGE 9870676.

References

M. P. Aylett. 1999. Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration. In *Proceedings of ICPhS–99, San Francisco*.

E. Charniak. 2001. A maximum-entropy-inspired parser. In *Proceedings of ACL–2001, Toulouse*.

J. T. Goodman. 2001. A bit of progress in language modeling. *Computer Speech and Language*, 15:403–434.

I. Kontoyiannis, P. H. Algoet, Yu. M. Suhov, and A.J. Wyner. 1998. Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inform. Theory*, 44:1319–1327, May.

I. Kontoyiannis. 1996. The complexity and entropy of literary styles. NSF Technical Report No. 97, Department of Statistics, Stanford University, June. [unpublished, can be found at the author's web page].

R. Kuhn and R. De Mori. 1990. A cache-based natural language model for speech reproduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6):570–583.

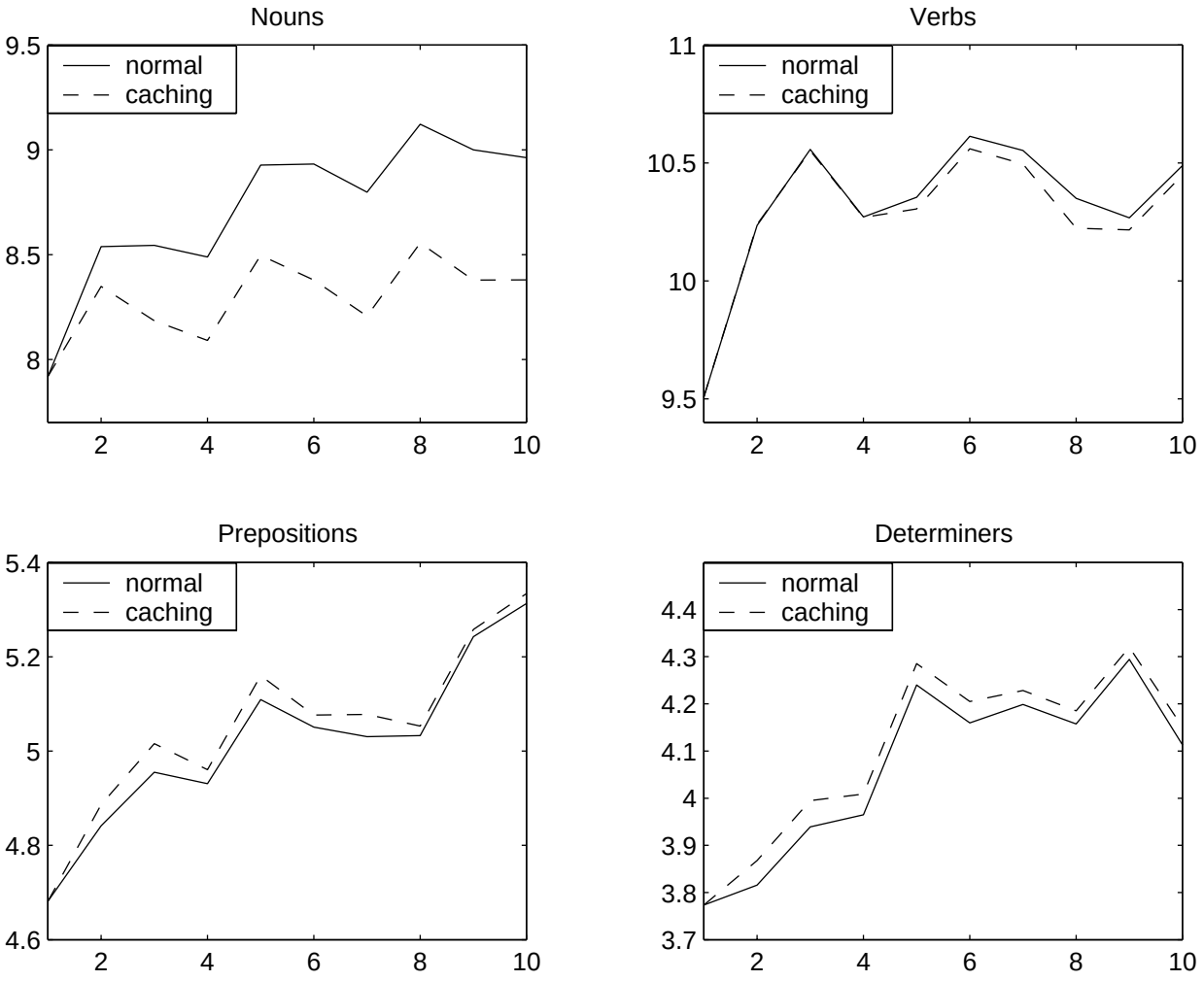

Figure 3: Comparing Parts of Speech

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics*, 19:313–330.

J. B. Plotkin and M. A. Nowak. 2000. Language evolution and information theory. *Journal of Theoretical Biology*, pages 147–159.

C. E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, July, October.