Compression without a common prior: an information-theoretic justification for ambiguity in language

Brendan Juba[1,2*] Adam Tauman Kalai[3] Sanjeev Khanna[4†] Madhu Sudan[3,5]
[1]School of Engineering and Applied Sciences, Harvard University
[2]Computer Science and Artificial Intelligence Laboratory, M.I.T.
[3]Microsoft Research New England
[4]Department of Computer and Information Science, University of Pennsylvania
[5]Department of Electrical Engineering and Computer Science, M.I.T.

Abstract: Compression is a fundamental goal of both human language and digital communication, yet natural language is very different from compression schemes employed by modern computers. We partly explain this difference using the fact that information theory generally assumes a common prior probability distribution shared by the encoder and decoder, whereas human communication has to be robust to the fact that a speaker and listener may have different prior beliefs about what a speaker may say. We model this information-theoretically using the following question: what type of compression scheme would be effective when the encoder and decoder have (boundedly) different prior probability distributions. The resulting compression scheme resembles natural language to a far greater extent than existing digital communication protocols. We also use information theory to justify why ambiguity is necessary for the purpose of compression.

Keywords: Compression, information theory, linguistics

1 Introduction

It is well-known that information theory sheds light on natural language in the following sense. Common words, such as "as" and "and" tend to be shorter than less frequent words such as "biomimicry." In this paper, we aim to strengthen the connection between information theory and the study of human communication. First, we point out that information theory justifies *ambiguity*, pervasive in natural language, by showing that it is necessary for efficient compression. Second, we design a compression scheme that bears a resemblance to natural language, to an extent well beyond that of existing compression and error-correcting schemes. Unlike standard compression schemes, it is *robust* to variations in the prior probability distribution between sender and receiver.

Natural language is ambiguous. One sentence could mean a variety of things in different contexts. At first thought, it is not clear that ambiguity serves any purpose, and communication may seem best when everything has the precision of mathematics with (ideally) exactly one interpretation. On such grounds, Wasow *et al.* (2005) call the existence of ambiguity in language surprising, and moreover, note that the relative *lack* of work or interest in the ambiguity of language by linguists is also surprising. Cohen (2006) discusses the various theories proposed for why language is ambiguous, but he concludes, "As far as I can see, the reason for the ambiguity of language remains a puzzle we simply don't know why language is ambiguous." According to Chomsky (2008), ambiguity illustrates that natural language was "poorly designed for communicative efficiency."

speculates that the primary purpose of ambiguity in language is not for succinct communication

[*]Supported by NSF award CCF-0939370.

[†]Supported in part by NSF Awards CCF-0635084 and IIS-0904314.

but for, "minimizing the complexity of rule systems."

However, it is easy to justify ambiguity to anyone who is familiar with information theory. Typical sentences, such as, *Alice said that Bob lied to Eve*, are ambiguous but shorter than clearer alternatives.[1] In *context*, the intended meaning is often clear, and hence shorter communication is preferred. This is exactly what information theory predicts – optimal compression is possible when there is a known prior probability distribution, p, over what is to be communicated. The common prior shared by a pair of communicating parties may be viewed as the shared context between them. The following manner of communicating would be essentially optimal in terms of minimizing expected communication length. For any natural number, n, a speaker who had in mind a certain thought would say n and mean *the nth most likely thought according to our shared prior distribution.*

Two problems with the above compression scheme stand out. First, it is very brittle in the sense that if the speaker and listener have even slightly different priors, every transmission may be completely erroneous. (This is true of Huffman coding as well.) Second, it clearly does not resemble human communication of any form. We show that these two problems are related by giving a compression scheme which is (a) robust to differences in priors, and (b) resembles human language.

1.1 The scheme and similarity to human disambiguation

We consider one-way (non-interactive) communication, in which there is a set of *messages*, representing what the sender would like to communicate (an idea, the true intended meaning of the communication). There is also a set of *encodings*, which represent the actual communication. For simplicity, we may think of the encoding as a single written sentence, but it could equally be an email, an elaborate hand gesture, or an utterance of arbitrary length. Some encodings are longer than others, and it is desirable to (a) ensure that the receivers recover the intended message, and (b) minimize the encoding length.

The sender has a prior probability distribution, p over messages. This prior distribution is determined by the context in which the discussion takes place, and to some extent the speaker's knowledge and all of her own experiences. The sender also chooses a parameter $\alpha \geq 1$ reflecting how broad an audience to whom her communication must be clear. For example, if the sender is writing a paper for people within her community, she would choose a smaller α then if she were writing for an interdisciplinary audience. The receiver has a potentially different prior distribution, q. The communication will be clear as long as q is within an α factor of p, i.e., $\frac{1}{\alpha}p(m) \leq q(m) \leq \alpha p(m)$ for all messages m.

Figure 1 depicts an underlying a bipartite graph between messages and encodings. This graph can be viewed as a *dictionary*: for each encoding it specifies a set of possible messages (meanings). It is assumed that this underlying graph (we postpone describing how it is chosen) is commonly known to both people and serves roughly the same purpose as a language. The receiver's decoding procedure is natural: given a received encoding, he chooses the *most likely* compatible message according to *his* distribution, i.e., the message most likely under q which has an edge to the received encoding. The sender, assuming that q is within a factor of α of p, chooses a minimal-length encoding that will guarantee correct decoding for all such q. This amounts to being the shortest encoding where the intended message has a significantly higher (α^2 factor) probability than any another possible interpretation.

The bipartite graph (i.e., dictionary) is chosen based upon some parameters. We give two instantiations. The first is simpler but has infinitely many parameters. The second is based upon universal hash functions and has parameters that require a number of bits which is logarithmic in the number of messages. This mirrors the Principles and Parameters Theory of linguistics (see, e.g., Chomsky and Lansik, 1993), which states that a small number of parameters characterize each language. In natural language, it would be infeasible to print a "sentence dictionary" of what every sentence or document might mean in any context. However,

[1] The sentence *Alice said, "Bob lied to Eve"* implies a direct quotation and therefore has a different meaning than the intended, *Alice said something to someone, and that something was that Bob lied to Eve*.

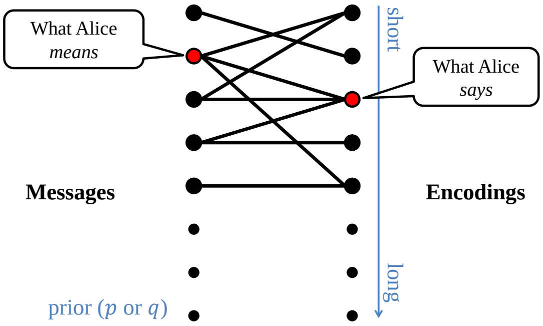

Figure 1: In our compression schemes, there is a common-knowledge "dictionary," a bipartite graph connecting encodings to possible messages. The messages, on the left side, are the possible ideas that the sender may wish to convey. The right side has each possible encoding, e.g., written sentences, longer units of text, or any other form of communication across a medium. Each person has a private prior distribution over messages. The bipartite graph has an edge between an encoding and each meaning that it might plausibly represent. In our scheme, the decoder simply chooses the *most likely* message that is adjacent to the transmitted encoding, according to the receivers prior. The sender chooses the shortest encoding which is guaranteed to be correctly decoded.

such a mapping is, to some extent, implicitly computable in people's mind. People would largely agree that the sentence *Alice said that Bob lied to Eve* could mean that *Alice said that (Bob lied to Eve)* or that *Alice told Eve that Bob lied*. Of course, there will never be perfect agreement on the complete set of possibilities, just as different dictionaries do not agree on definitions or even the set of legal words. In this example, some may argue that, in the above, Alice might be indicating that Bob was lying (on a bed, perhaps), and avoided the grammatically correct version: *Alice said that Bob lay to Eve*. While there will always be gray areas and exceptions to most linguistic rules, to a first approximation this dictionary model of language is more faithful than either of the two extremes: each sentence has exactly one interpretation (like an ideal computer programming language), or any sentence can mean anything in the right context (as in compression schemes such as Huffman coding).

We feel that this procedure also resembles human language both in terms of listening and speaking, or at least to a greater extent than existing compression schemes. In terms of disambiguation, it seems natural for a listener to take the most likely plausible interpretation in the "dictionary," under his prior over what he expects the speaker to mean. Conversely, it is the speaker's duty to communicate in such a manner that any listener in her audience will believe that the intended message is the most likely interpretation of what is said. And of course it is ambiguous – a certain encoding may be decoded differently depending on the decoder's prior (context). Furthermore, these properties arise naturally out of a mathematical goal of provable efficiency in encoding length.

Finally, we also mention a technique whereby one can reduce the dictionary size. This *pruning* step only leads to a slight improvement in efficiency. However it also resembles an effect that occurs in language. It takes advantage of the fact that a speaker would not normally use an unnecessarily complicated expression for a simple idea that could be described in a shorter unambiguous fashion. In mathematics, consider the two definitions,

$$H(p) \doteq \sum_x p(x) \log 1/p(x)$$

$$f(x) \doteq \log x/x$$

Here, mathematically sophisticated readers will naturally interpret $\log 1/p(x)$ as meaning

$\log(1/p(x))$ rather than $\log(1)/p(x) = 0$. On the other hand, the $\log x/x$ will be $\log(x)/x$ rather than $\log(x/x) = 0$. In both cases, the listener is performing higher-order reasoning. In particular, the listener would have expected a simpler, unambiguous definition, like $f(x) \doteq 0$, if the intended meaning were 0. While the savings in communication is modest, such short-cuts are regularly used by mathematicians, who generally have a strong desire to avoid ambiguity. For an English example, consider the example of sentence, *You may step forward when your number is called.* The implication is that you may *not* step forward *before* your number is called, for if that was not the intention, the sentence *You may step forward at any time* could have been used.

Such instances where listeners use higher-order reasoning to determine a meaning of an utterance beyond what the utterance literally suggests were first studied by Grice (1975), who called this process "conversational implicature." In Grice's theory, he put forward the *cooperative principle* that supposed that speakers adhered to a list of maxims – including, "Make your contribution as informative as required" and "Be brief (avoid unneccessary prolixity)," among many others – and he argued that listeners will logically infer the speaker's true meaning by taking the speaker's adherence to these maxims as axioms. Grice's maxims were subsequently reformulated into a few more coherent principles by numerous authors (Levinson (2000) gives a nice summary). Our model suggests a simpler alternative account of many instances of conversational implicature: the speaker simply says as little as possible to overcome the disagreement with the listener's prior, trusting the listener to reason that any other (unintended) likely meanings would have had shorter expressions, e.g., as done by our second decoding scheme.

Other authors have noticed that conversational implicature might arise from the desire to communicate more efficiently—Sperber and Wilson (1995) in particular dwell on this point; conversely, many authors also noticed that conversational implicatures might be closely related to ambiguity, specifically that they might exist for similar reasons and employ similar mechanisms. Indeed, for all the bitter disagreements that appear to exist between Sperber and Wilson and Levinson (2000), they strongly agree on these points. The difference in our work is that while on the one hand we make no promises about being able to account for vast ranges of phenomena like Levinson or Sperber and Wilson do, on the other hand we show that effects like conversational implicature can arise from surprisingly minimial and uncontroversial considerations. Indeed, our model is consistent with the premises laid out by Sperber and Wilson prior to the point where they begin speculating about cognitive architectures, and is arguably "more obvious" (in hindsight) than the model they end up with.[2]

1.2 Interpretation and applications

Designing and recognizing the similarity between nature and engineering informs our understanding of both. Consider, for example, the striking similarity between the camera and human eye. These similarities suggest that certain aspects of the eye are not artifacts of poor evolution, but instead may serve a purpose. In the same way as connections between photography and human vision deepen our understanding, we hope that robust compression schemes may help connect information theory and the study of human communication.

Second, there may be situations where two computer systems need to communicate in a compressed fashion, but they do not share exactly the same prior. Consider, for example, a computer compressing a document to be sent to a printers. Now, a fixed compression scheme could be agreed upon in advance. However, for compatibility reasons, this compression scheme would remain fixed for many years, and it may become poorly suited for a certain category of documents that emerge years later. For example, if many people started printing many documents with the same fixed logo on it, the computers and printers may adapt.

[2] Although the model presented by Sperber and Wilson (1995) is rather informal, the formalizations based on information theory presented by, e.g., Blutner (1998), and formalizations based on game theory presented by, e.g., Parikh (1992), Merin (1997), and van Rooy (2001) naturally end up being on the one hand more intricate, but again, on the other hand are intended to deal with a wider range of effects, and therefore generally incomparable.

The idea here is that computers and printers could *learn* and periodically update their priors based on the documents they transmit, so that they may continue to compress well under changing environment. The following modification here may be useful. Suppose there is a simple way to verify if the correct document was reconstructed, which may be achieved by a checksum or more elaborate mechanism. Then notice that the parameter α can be tuned adaptively: communication with a smaller α may be attempted first, and if that fails, a retry with a larger α may be used, and so forth. Such a system would be adaptive in the sense that, years down the road, any computer and printer employing this protocol could communicate succinctly, even if they had never previously encountered each other, with a logarithmic overhead in terms of how different the documents they had seen were. This type of copying nature for engineering purposes has been recently popularized under the term *biomimicry*.

1.3 Related work

In recent independent work, Piantadosi *et al* (2010) justify ambiguity in natural language as we do by an information-theoretic argument, but do not enter into the realm of different priors. A similar technical question about compressing with different priors, arises in recent independent work by Braverman and Rao (2010). The focus of their work is attaining optimal bounds for reducing interactive communication complexity, rather than modeling human communication. A related notion, the quantity *relative entropy*, answers the following question. When two parties communicate using a protocol designed for common prior q, how long will messages be when the encoder actually chooses them from p? In this case, the encoder must know q exactly, which is unrealistic in many settings.

Universal compression schemes, such as the Lempel-Ziv (1978) scheme, compress without knowledge or dependence on a prior, so it is universal for all sources. Asymptotically optimal compression is guaranteed for *ergodic* sources, e.g., those generated by small state Markov chain. However, any such prior-free encoding will fail to take advantage of the rich shared knowledge base that enables two parties to communicate a significant amount of information in a short document or even a single sentence. In short, existing compression schemes, including Huffman, Lempel-Ziv, or algebraic codes that we have not described, are clearly poorly suited for human communication.

Other prior work has also explored communication in the setting where the sender and receiver are somehow different. For instance, Juba and Sudan (2008) and Goldreich, Juba and Sudan (2009) considered how *interacting* pairs may achieve certain goals that can be achieved only by communication. Our work, while inspired by such work, is different in several aspects: It focusses on a different objective, namely to reduce the number of bits used to communicate the message. Also, we focus on the *non-interactive* setting, and the quantitative bit-efficiency of our protocol is central to our quest. Finally, our goal is to capture phenomena that may explain some of the apparent artifacts of *natural language*.

2 Formal model

There is a set of expressions which we denote by X, and a set of meanings which we denote by M. We assume that M is finite or countably infinite and, for clarity, we take $X = \{0,1\}^*$.[3] A *context* provides a probability distribution over meanings, and $\Delta(M)$ denotes the set of probability distributions over M.

We assume that the encoding scheme and decoding scheme may share a common parameter $\theta \in \Theta$, chosen from some probability distribution μ. In our schemes, this parameter corresponds to the aforementioned bipartite graph. An encoder is a function, $\mathcal{E} : M \times \Delta(M) \times \Theta \to X$, written $\mathcal{E}_\theta(m, p)$, from meanings to expressions. Similarly, a decoder $\mathcal{D} : X \times \Delta(M) \times \Theta \to M$, written $\mathcal{D}_\theta(m, p)$, is a function from expressions and contexts to meanings. Note that the parameter θ is chosen without regard to p or q. When θ is clear from context, we will write $\mathcal{E}(m, p)$ and $\mathcal{D}(x, q)$. A *randomized compression scheme* is a sextuple $(X, M, \Theta, \mathcal{E}, \mathcal{D}, \mu)$, where μ is a probability distribution over Θ. Two probability distributions, $p, q \in \Delta(M)$ are called *α-close*, for *ambiguity parameter* $\alpha \geq 1$, if $p(m) \leq \alpha q(m)$ and $q(m) \leq \alpha p(m)$ for all $m \in M$.

[3] While we recognize that set of all finite binary strings is clearly different than the richly-structured sets used in real language, our choice of X will suffice to make our main points.

Definition 1. *A randomized compression scheme is called α-*robust *if for any α-close pair, (p, q) and any $m \in M$, $\Pr_{\theta\sim\mu}[\mathcal{D}_\theta(\mathcal{E}_\theta(m, p), q) = m] = 1$. The* entropy *of the scheme (on p) is defined to be* $\mathrm{E}_{\theta\sim\mu,m\sim p}[|\mathcal{E}_\theta(m, p)|]$.

Note that it typically suffices to describe a decoding procedure $\mathcal{D}(x, q)$ since the optimal matching compression function $\mathcal{E}(m, p)$ simply selects the shortest string x such that $\mathcal{D}(x, q) = m$ for all q that are α-close to p. (Recall that the encoder is assumed to know α in advance.)

3 Our compression scheme

In this compression scheme, we assume that the encoder and decoder share a common *infinite* parameter sequence $\langle r_m^{(i)}\rangle_{i=1}^\infty$, where $r_m^{(i)} \in \{0, 1\}^i$ for each $m \in M$, chosen uniformly at random and independently. In other words, for each message and each length $i = 1, 2, \ldots$, an independent random binary string of length i is chosen and shared between the encoder and decoder.[4] This determines a bipartite graph between messages and $\{0, 1\}^*$ by connecting each message m to $r_m^{(i)}$, for each i. As mentioned, this is similar to a dictionary. In section 6, we give a more practical scheme that requires a number of random bits that is logarithmic in the number of messages.

On encoding x of length $i = |x|$, the decoder chooses the *most likely* message m (that of greatest $q(m)$) among those messages such that $r_m^{(i)} = x$. Formally, the scheme is as follows.

Compression scheme. The encoding algorithm and decoding algorithm share randomness, namely infinite sequences of random strings $\theta = \langle r_m^{(i)}\rangle_{i=1}^\infty$.
To encode $m \in M$:

- Send $r_m^{(i)}$ where i is the smallest natural number such that: $p(m) > \alpha^2 p(m')$ for all messages m' where $r_m^{(i)} = r_{m'}^{(i)}$.[5]

To decode $x \in \{0, 1\}^*$:

- Let $i = |x|$ and $S = \{m \in M \mid r_m^{(i)} = x\}$. Output $\arg\max_{m\in S} q(m)$.[6]

Observation 1. *For any $\alpha > 1$, and uniformly random $r_m^{(i)} \in \{0, 1\}^i$, the compression scheme is α-robust. For any $p \in \Delta(M)$, its entropy is at most $H(p) + 2\lg(\alpha) + 2$.*

In the above, $H(p)$ is the standard entropy of probability distribution p, defined by $\sum_m p(m) \lg 1/p(m)$.

Proof. The correctness of decoding follows from the fact that for any α-close p and q, if $p(m) > \alpha^2 p(m')$ then $q(m) \geq p(m)/\alpha > \alpha p(m') \geq q(m')$. With probability 1, there will be such an i that $p(m) > \alpha^2 p(m')$ for all messages m' where $r_m^{(i)} = r_{m'}^{(i)}$. So with probability 1, the message that was encoded is necessarily the most likely $m \in S$ for decoding.

It suffices to show that the expected encoding length of a message m is at most $\lg(\alpha^2/p(m)) + 2$. To see this, note that there are less than $\alpha^2/p(m)$ messages, different from m, with probability at least $p(m)/\alpha^2$. Call this set T and let $s = |T| < \alpha^2/p(m)$. Consider the probability that any other message $m' \in T$ collides with m on the i-bit encoding ($r_m^{(i)} = r_{m'}^{(i)}$). For $i = \lceil\lg(s)\rceil + k$, by the union bound, this probability is at most $s2^{-(\lg(s)+k)} \leq 2^{-k}$. Using the fact that for any nonnegative integer random variable V, $\mathrm{E}[V] = \sum_{i=1}^\infty \Pr[V \geq i]$, we have that the expected number of bits in common is at most $\lceil\lg(s)\rceil + \sum_{k=0}^\infty 2^{-k} \leq \lg(s) + 2$.

□

It is not difficult to show that no α-robust scheme can achieve entropy better than $H(p) + \lg(\alpha)$ for all p. On the other hand, we show below that there exist distributions for which the entropy bound achieved by our scheme is $H(p)+(2-o(1))\lg(\alpha)$ (i.e. our analysis is essentially tight).

Claim 1. *For any $\varepsilon \in (0, 1)$, there exists a distribution p and an $\alpha = \alpha(\varepsilon)$ such that the entropy of the above compression scheme is at least $H(p) + (2 - \varepsilon)\lg\alpha$.*

[4] For simplicity, the algorithms are described using infinitely many random bits. More practical versions are possible.

[5] In the (zero probability) event where there is no such number, send 0.

[6] To formally define $\mathcal{D}$, we must define how the decoding scheme behaves if there is not a unique maximum (or $S = \emptyset$). In this case, we could designate a fixed message m_0 and output that message.

Proof. Fix $k = \lceil 3/\varepsilon \rceil$, and $\alpha = 2^{k^2}$. Now consider a distribution p defined as follows: for each $i \in \{1, 2, \ldots, k\}$, the distribution p contains α^{2i} messages that each have probability $1/k\alpha^{2i}$. Then

$$H(p) = \frac{1}{k}\left(\sum_{i=1}^{k} \lg(k\alpha^{2i})\right) = (k+1)\lg\alpha + \lg k.$$

On the other hand, the entropy of the compression scheme is bounded from below by,

$$\frac{1}{k}\left(\sum_{i=1}^{k-1} \lg(\alpha^{2i+2}) + \lg(\alpha^{2k})\right) = \\ (k+1)\lg\alpha + 2\lg\alpha - \frac{2\lg\alpha}{k}.$$

Since $(2\lg\alpha)/k + \lg k \leq \varepsilon \lg\alpha$ for our choice k and α, the claim follows. □

An interesting question is if there is a compression scheme that matches the $H(p) + \lg(\alpha)$ bound.

4 The need for ambiguity

In this section, we show that any unambiguous compression scheme requires many bits to communicate. This holds even for nonrobust communication, i.e., for α=1. Formally, say an encoder is *unambiguous* if for all $\theta \in \Theta$, $m, m' \in M$, and $p, p' \in \Delta(M)$, if $\mathcal{E}_\theta(m, p) = \mathcal{E}_\theta(m', p')$ then $m = m'$. Define the dirac probability distribution δ_m by $\delta_m(m) = 1$ and $\delta_m(m') = 0$ for $m' \neq m$.

Observation 2. *For any unambiguous encoder on finite message set M, there is a message m such that δ_m has expected entropy of* $\lg(|M|) - 1$.

Hence, the trivial encoding scheme of encoding each message by a unique length-$\lg M$ binary string, independent of p, is essentially optimal even for probability distributions δ_m, where $H(\delta_m) = 0$.

Proof. Note that for any θ, the function $f(m) = \mathcal{E}_\theta(m, \delta_m)$ is injective. Hence, by a standard counting argument, for any fixed θ, $\mathrm{E}_{m \in_{\mathcal{U}} M}[\ |\mathcal{E}_\theta(m, \delta_m)|\ ] \geq \lg(|M|) - 1$, where the expectation of is taken over uniformly random $m \in M$. Thus

$$\mathrm{E}_{m \in_{\mathcal{U}} M, \theta \sim \mu}[\ |\mathcal{E}_\theta(m, \delta_m)|\ ] \geq \lg(|M|) - 1.$$

Hence, there exists some message m such that $\mathrm{E}_{\theta \sim \mu}[\ |\mathcal{E}_\theta(m, \delta_m)|\ ] \geq \lg(|M|) - 1$, as is claimed. □

5 Higher-order disambiguation and pruning the dictionary

If a message m has a unique encoding of length i, then it seems unnecessary to disambiguate between m and other messages on encodings of length greater than i. This idea can be used to decrease the number of edges in the bipartite graph as well as average number of bits communicated. Given parameter vector $\theta = \langle r_m^{(i)} \rangle_{i=1}^{\infty}$, where $r_m^{(i)} \in \{0,1\}^i$ for each $m \in M$, we choose the following pruned vector $\hat{\theta} = \langle \hat{r}_m^{(i)} \rangle_{i=1}^{\infty}$, constructed as follows. Define $M_1 = M$ and,

$$M_{i+1} = \{m \in M_i \mid \exists m' \in M_i \text{ s.t. } m' \neq m \\ \text{and } r_m^{(i)} = r_{m'}^{(i)}\}.$$

M_i are the set of messages that do not have a completely unambiguous encoding of length less than i. Finally, for each m and i, set,

$$\hat{r}_m^{(i)} = \begin{cases} r_m^{(i)} & \text{if } m \in M_i \\ -1 & \text{otherwise.} \end{cases}$$

In other words, a message which has a unambiguous encoding of length i will be not have any encodings of greater length.

Observation 3. *For any $\alpha > 1$, our compression scheme using $\hat{r}$ instead of r is α-robust and has entropy no greater than the entropy when using r. There are probability distributions p for which it has strictly lower entropy.*

As can be seen from the proof below, "most" nontrivial probability distributions will have strictly lower entropy in the higher order scheme.

Proof. The proof of α-robustness is exactly the same as in the first case. Clearly, the encoding of any message cannot be longer than that of the second compression scheme, if the two share the same random strings. Finally, take three messages $M = \{a, b, c\}$ and $p(a) = p(b) = p(c) = 1/3$. With positive probability, $r_a^{(1)} = 0$, $r_b^{(1)} = r_c^{(1)} = 1$, $r_a^{(2)} = r_b^{(2)} = 00$, and $r_c^{(2)} = 01$. In this case, the compression scheme encodes b by a string of length greater than 2 while the higher-order scheme encode b by 00. □

6 Using fewer random bits

As stated, our compression scheme requires *infinite* randomness, for finite message spaces, M. We now give a variation with $O(\log(|M|))$ random bits, using Universal Hash Functions (Carter, and Wegman, 1979). Again, we do not change the compression scheme but simply the *dictionary*, i.e., we apply the compression scheme described earlier with a different $\langle r_m^{(i)} \rangle$.

We assume w.l.o.g. that each message corresponds to an ℓ-bit string where $\ell = \lceil \log(M+1) \rceil$, that is, $m \in \{0,1\}^\ell$. Let π be any prime in the interval $[2^\ell, 2^{\ell+1})$ (it exists by Bertrand's postulate). The language will have a pair a, b of random parameters where $a \in \{1, 2, \ldots, \pi - 1\}$, and $b \in \{0, 1, 2, \ldots, \pi - 1\}$. We view each message m as an integer in $\{0, 1, \ldots, 2^\ell - 1\}$, and define,

$$r_m^{(i)} = \begin{cases} \big((am+b)(\text{mod } \pi)\big)(\text{mod } 2^i) & \text{if } i \le \ell \\ m & \text{if } i = \ell + 1 \\ -1 & \text{if } i > \ell + 1 \end{cases}$$

Observation 4. *Let $a \in \{1, 2, \ldots, \pi - 1\}$ and $b \in \{0, 1, 2, \ldots, \pi - 1\}$ be chosen uniformly at random. Then the compression scheme with r as defined above is α-robust and, for any $p \in \Delta(M)$, has expected entropy $H(p) + 2 \lg \alpha + 2$.*

Proof. By construction, every message has a different $r_m^{(\ell+1)}$, hence no message will require more than $\ell + 1$ bits to encode and the protocol is α-robust. It suffices to show that the expected encoding length of a message m is at most $\lg(\alpha^2/p(m)) + 2$.

To see this, again note that there are less than $\alpha^2/p(m)$ messages, different from m, that have probability at least $p(m)/\alpha^2$. Call this set T and let $s = |T| < \alpha^2/p(m)$. Notice that a fixed $m' \in T$ collides with m on the i-bit encoding, when $r_m^{(i)} = r_{m'}^{(i)}$ which happens iff $(am + b) \pmod \pi$ and $(am' + b) \pmod \pi$ agree on the last i bits. So the collisions are correlated, and in particular any two messages will collide on $r_m^{(i)}$ for $i = 1, 2, \ldots,$ up to the number of trailing bits that they agree on. A simple and standard argument shows that for any pair of distinct messages m, m', as a ranges over $\{1, 2, \ldots, \pi - 1\}$ and b ranges over $\{0, 1, 2, \ldots, \pi - 1\}$, we have (i) $(am+b) \pmod \pi \neq (am'+b) \pmod \pi$, and (ii) $((am + b) \pmod \pi, (am' + b) \pmod \pi)$ range over all possible $\pi(\pi - 1)$ pairs of values. Thus over random choice of a, b, for any $i \le \ell + 1$, we have that $\Pr[r_m^{(i)} = r_{m'}^{(i)}] \le 2^{-i}$. Hence for $i = \lceil \lg(s) \rceil + k$, by the union bound, the probability that any message in T agrees with m on i bits is at most $s\big(2^{-(\lg(s)+k)}\big) \le 2^{-k}$. As before, using the fact that for any nonnegative integer random variable $V \in \{1, 2, \ldots, \ell+1\}$, $\mathrm{E}[V] = \sum_{i=1}^{\ell+1} \Pr[V \ge i]$, we have that the expected encoding length is at most $\lceil \lg(s) \rceil + \sum_{k=0}^{\ell} \big(2^{-k}\big) \le \lg(s) + 2$.

□

Thus $O(\log |M|)$ random bits suffice.

7 Conclusions

We have shown that ambiguity is necessary for compression, and that a natural variation on Shannon-type of communication leads to robust compression schemes that are more similar to how humans communicate. The case of $\alpha = 1$ corresponds to classical compression with a common prior. In this case, for Shannon's fundamental question of how many bits are required to compress a message from a single distribution, the beautiful answer is Huffman coding (Huffman, 1952). However Huffman coding is not robust to different priors. It is not even clear what metric should be used to judge optimality with respect to robust compression.

Second, our model is unrealistic in many ways. For example, the encoder must choose a single α and is required to be precise to all α-close priors. In some cases, an encoder may consider some misinterpretations to be more "costly" than others, i.e., there may be a cost function $c : M \times M \to \mathbb{R}_+$ ($c(m, m')$ is the cost of interpreting message m to be m', and $c(m, m) = 0$), and an encoder choosing amongst ambiguities may wish to avoid certain mistakes. For example, the sentence, *he is a* *tireless student and brilliant researcher,* could potentially mean *he is a tireless student, and he is a great researcher* or *he is a tireless student, and he is a tireless great researcher*, but a confusion would not be serious. On the other hand, the sentence, *you would be lucky to get him to work for you* is ambiguous and the difference in meaning is very important.

Finally, we have not considered computational efficiency. Day to day, it does not seem that computational limitations are the cause of most failures to communicate. However, there are some sentences that are notoriously difficult to parse, called *garden path sentences*, such as the classic sentence, *The horse raced past the barn fell.* Similarly, riddles are computationally challenging to solve. It would be very interesting to design computationally-efficient robust compression schemes.

Acknowledgments. We are grateful to Noam Chomsky, Ted Gibson, and Ghid Maatouk, for helpful discussions.

References

[1] Blutner, Reinhard. “Lexical Pragmatics,” *J. Semantics* 15:115-162, 1998.

[2] Braverman, Mark and Anup Rao. “Efficient Communication Using Partial Information.” *Electronic Colloquium on Computational Complexity* (ECCC) TR10-083, 2010.

[3] Carter, Larry and Mark N. Wegman. “Universal Classes of Hash Functions.” *Journal of Computer and System Sciences* 18 (2): 143-154, 1979.

[4] Chomsky, Noam. “On phases.” In *Foundational Issues in Linguistic Theory. Essays in Honor of Jean-Roger Vergnaud*, C. Otero et al. (eds.), 134-166. Cambridge, MA: The MIT Press, 2008.

[5] Chomsky, Noam and Howard Lasnik. *Principles and Parameters Theory, in Syntax: An International Handbook of Contemporary Research*. Berlin: de Gruyter, 1993.

[6] Cohen, Ariel. “Why ambiguity?” In *Between 40 and 60 Puzzles for Manfred Krifka*, Hans-Martin Gaertner, Sigrid Beck, Regine Eckardt, Renate Musan, and Barbara Stiebels (eds.), 2006.

[7] Goldreich, Oded, Brendan Juba, and Madhu Sudan. “A Theory of Goal-Oriented Communication.” *Electronic Colloquium on Computational Complexity* (ECCC) TR09-075, 2009.

[8] Grice, H. Paul. “Logic and Conversation,” In *Syntax and Semantics 3: Speech Acts,* P. Cole & J. L. Morgan (eds.), Academic Press, New York. 1975.

[9] Huffman, David. “A Method for the Construction of Minimum-Redundancy Codes,” *Proceedings of the I.R.E.*, pp. 1098-1102, 1952.

[10] Juba, Brendan and Madhu Sudan. “Universal Semantic Communication I.” In *Proceedings of 40th Annual ACM Symposium on Theory of Computing* (STOC), pp. 123-132, 2008.

[11] Lempel, Abraham, and Jacob Ziv. “Compression of individual sequences via variable-rate coding.” In *In Proceedings of IEEE Transactions on Information Theory*, pp. 530-536, 1978.

[12] Levinson, Stephen C. *Presumptive Meanings.* MIT Press, Cambridge, MA. 2000.

[13] Merin, Arthur. “Information, relevance, and social decisionmaking,” In *Logic, Language, and Computation,* Vol. 2, L. Moss, J. Ginzburg, & M. De Rijke (eds.), Stanford. 1997.

[14] Parikh, Prasant. “A game-theoretic account of implicature,” In *Proceedings of 4th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pp.85-94, 1992.

[15] Piantadosi, Steven, Harry Tily, and Edward Gibson. The communicative function of ambiguity in language. Manuscript, 2010.

[16] Shannon, Claude. “A Mathematical Theory of Communication,” *Bell Sys. Tech. J.* (27):379-423,623-656, 1948.

[17] Shannon, Cluade. “Prediction and Entropy of Printed English,” *Bell Sys. Tech. J.* (3):50-64, 1950.

[18] Sperber, Dan and Deirdre Wilson. *Relevance,* second edition. Blackwell, Cambridge, MA. 1995.

[19] van Rooy, Robert. “Relevance of communicative acts.” In *Proceedings of 8th Conference on Theoretical Aspects of Rationality and Knowldege (TARK)*, pp.84-96, 2001.

[20] Wasow, Thomas, Amy Perfors, and David Beaver. “The Puzzle of Ambiguity.” In *Morphology and the Web of Grammar: Essays in Memory of Steven G. Lapointe,* O. Orgun and P. Sells (eds.), CSLI Publications, Stanford. 2005.