SANDIA REPORT

SAND2009-8168
Unlimited Release
Printed December 2009

Information and Meaning Revisiting Shannon's Theory of Communication and Extending it to Address Today's Technical Problems

Travis L. Bauer

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2009-8168
Unlimited Release
Printed December 2009

Information and Meaning
Revisiting Shannon's Theory of Communication and Extending it to Address Today's Technical Problems

Travis L. Bauer
Analytics and Cryptography
Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185-1235
tlbauer@sandia.gov

Abstract

This paper has three goals. The first is to review Shannon's theory of information and the subsequent advances leading to today's statistics-based text analysis algorithms, showing that the semantics of the text is neglected. The second goal is to propose an extension of Shannon's original model that can take into account semantics, where the "semantics" of a message is understood in terms of the intended or actual changes on the recipient of a message. The third goal is to propose several lines of research that naturally fall out of the proposed model.

Acknowledgment

I'd like to thank Randall Laviolette and the 5600 research ideas group for the opportunity to present and discuss an early version of this idea.

I'd also like to thank the Networks Grand Challenge (and Philip Kegelmeyer, the PI) for providing a small amount of funding to work on this paper.

Contents

Appendix

List of Figures

Preface

Introduction

Each computational approach to solving some problem rests on an underlying model or set of models that describe how key phenomena in the real world are represented and how they are manipulated. These models are both liberating and constraining. They are liberating in that they suggest a path of development for new tools and algorithms. They are constraining in that they intentionally ignore other potential paths of development.

Modern statistical-based text analysis algorithms have a specific intellectual history and set of underlying models rooted in Shannon's theory of communication. For Shannon, language is treated as a stochastic generator of symbol sequences. Shannon himself, subsequently Weaver, and at least one of his predecessors [1] are all explicit in their decision to exclude semantics from their models. This rejection of semantics as "irrelevant to the engineering problem" [11] is elegant and combined with developments particularly by Salton and subsequently by Latent Semantic Analysis, has led to a whole collection of powerful algorithms and an industry for data mining technologies.

However, the kinds of problems currently facing us go beyond what can be accounted for by this stochastic model. Today's problems increasingly focus on the semantics of specific pieces of information. And although progress is being made with the old models, it seems natural to develop or extend information theory to account for semantics. By developing such theory, we can improve the quality of the next generation analytical tools. Far from being a mere intellectual curiosity, a new theory can provide the means for us to take into account information that has been to date ignored by the algorithms and technologies we develop.

This paper will begin with an examination of Shannon's theory of communication, discussing the contributions and the limitations of the theory and how that theory gets expanded into today's statistical text analysis algorithms. Next, we will expand Shannon's model. We'll suggest a transactional definition of semantics that focuses on the intended and actual change that messages are intended to have on the recipient. Finally, we will examine implications of the model for algorithm development.

[1] Hartley, who Shannon said had an important influence on his life [4]

Motivating issues

What motivates this need to revisit Shannon's theory is that the problems we are trying to address using advanced analytic tools differ fundamentally from the kinds of problems today's algorithms were intended to address. In the past 40 years, there has been significant success at building tools for retrieving and categorizing documents and these tools have been built using algorithms that assume Shannon's theory of communication.

That's not to say that Shannon's theory is flawed. Semantics *were* irrelevant to the engineering problem Shannon was trying to address, i.e. the transmission of messages from point A to point B. Also, it turns out that perturbations in the underlying statistics of natural language is useful in differentiating documents from one another, so subsequent development of the ideas was successful applications without an underlying understanding of the meaning of the texts.

However, there is a growing class of problems that current techniques have not been as successful at solving and for which the underlying theory is ill suited. Shannon himself objected to getting on the "scientific bandwagon" of the over application of his theory to the semantics of communication [10], a bandwagon that may have started with Weaver's 1949 paper. Today's problems are different from the ones that Shannon et. al. worked on and push past the point where the theory is applicable. Today's advanced analytics tools are expected to operate based on the meaning of the data and are intended to compute more than relevance probability. Technology consumers are asking for systems that suggest "actionable intelligence" and not just identify documents that, if read manually by a knowledgeable expert, might be used to create actionable intelligence. The phrase "needle in the haystack" makes the problem sound easier than it really is. Needles are fundamentally different from hay in their composition and finding such glaring differences, even if small, are easy for today's algorithms. But the problems being addressed today don't require the ability to find something that is different from the surrounding information. It requires the ability to understand the semantics of the underlying information.

Tools that understand the meaning of the underlying data, if successful, must necessarily be built on a theory that extends the theory of information presented by Shannon. And it has to expand on his theory in a way that incorporates semantics.

This page intentionally left blank.

Chapter 1

Overview of Shannon's Theory of Communication

The Basic Model

It is hard to overstate the value of Shannon's theory of communication and the impact it has had on analytic algorithms to date. The paper is cited in academic papers in fields as diverse as text analysis, network communications, bio-informatics, and environmental conservation.

In his landmark paper [11], Claude Shannon described a model and a way to quantify information. The key problem, as he explained it, was one of communication, i.e. how to produce at some point a message selected at some other point. His critical contribution was to measure the amount of information in the message without needing to understand that message's meaning. He accomplished this by jettisoning any attempt to deal with the meanings of messages and to treat a language as a stochastic generator of symbol sequences.

Figure 1.1 shows Shannon's diagram of a general communication system. In his paper, he was concerned with taking a message generated at the Information Source and reproducing that message at the Destination. These messages could take the form both of continuous functions or of discrete sets of symbols.

Explicit decision not to deal with semantics

When we go back to the theory underlying these methods, we find repeated warnings that effectively tell us that the theory does not address semantics. Hartley, in the 1920's, as well as Shannon and Weaver, in the 1940's, all repeatedly assert that the methods of measuring information that they were developing are not analyses of the meanings of the documents. Shannon states in the abstract of the 1948 paper that (emphasis is mine)

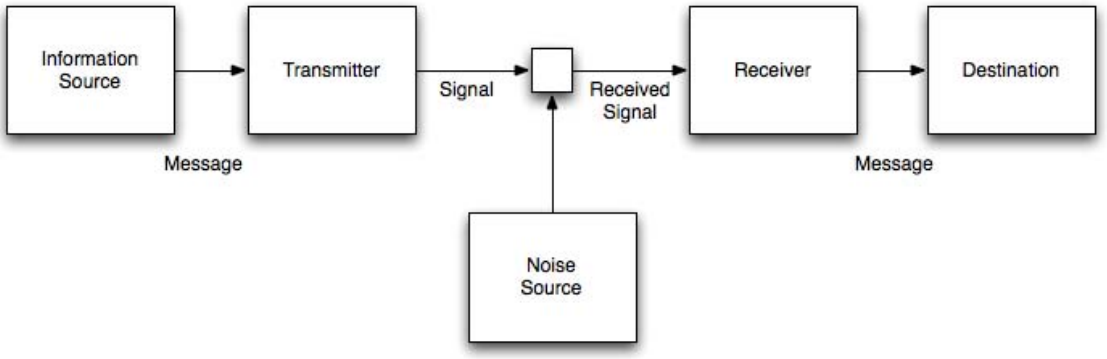

Figure 1.1. Shannon's schematic diagram of a general communication system

> Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. *These semantic aspects of communication are irrelevant to the engineering problem.* [11]

This echos Hartley from 20 years previously (the first part of the quotation below is the opening sentence of the paper) (emphasis is mine):

> A quantitative measure of information is developed which is *based on physical as contrasted with psychological considerations* . . . Hence estimating the capacity of the physical system to transmit information we should ignore the question of interpretation . . . By this means the psychological factors and their variations are eliminated and it becomes possible to set up a definite quantitative measure of information based on physical considerations alone. [5]

A close examination of the paper will confirm that this decision not to address semantics is not an aside or secondary consideration for the authors, but critical to the success of the methods they develop. Much of the ideas derived in their papers could not be derived unless they made this decision to ignore semantics.

Weaver's warnings

The year after Shannon's landmark paper, a number of questions appear to have arisen, including questions regarding the relationship between information and meaning, that gave rise to the need for a clarifying paper. Weaver clarified a number of things, including the relationship between information and semantics (which he calls meaning). Significantly, he affirms and expands on Shannon's assertion that information and meaning are two very different things. Weaver writes:

> "The concept of information developed in this theory at first seems disappointing and bizarre disappointing because it has nothing to do with meaning, and bizarre because it deals not with a single message but rather with the statistical character of a whole ensemble of messages, bizarre also because in these statistical terms the two words information and uncertainty find themselves to be partners." [12]

This point is made in the same paper more than once and in different ways.

> "An engineering communication theory is just like a very proper and discreet girl accepting your telegram. She pays no attention to the meaning . . ."

> "The word information, in this theory, is used in a special sense that must not be confused with its ordinary usage. In particular, information must not be confused with meaning."

> "One has the vague feeling that information and meaning may prove to be something like a pair of canonically conjugate variables in quantum theory, they being subject to some joint restriction that condemns a person to the sacrifice of the one as he insists on having much of the other."

Weaver's final dismissal of meaning is to fold semantics into the model as just another kind of symbol that could be transmitted. Whatever one might mean by meaning, Weaver was confident that it could be expressed as a message in some vocabulary. And having been expressed as a message, it falls under Shannon's model.

This page intentionally left blank.

Chapter 2

History of Intellectual Development After Shannon

So far, we've discussed the distinction between meaning and information within Shannon's framework. The next issue is to show that subsequent algorithm developments did not re-introduce semantics. This fact is initially unintuitive. In spite of the decision not to address semantics, the algorithms derived from there have had success at processing documents according to their meaning, even if on limited, focused problems. Additionally, core Latent Semantic Analysis (LSA) researchers have gone so far as to propose that LSA as a psychologically plausible theory of meaning [7].

Overview of the History

The path from Shannon's Theory of Communication to text analysis technology today goes through several major advances, the most significant of which are the development of the bag-of-words document/term matrix, generally associated with Salton, and the development of Latent Semantic Analysis, associated with Landauer and Dumais.

From Shannon to Salton

We can see in the literature two intellectual advances that led to the use of linear algebra techniques for statistical text analysis. The first is the association of utility with the probability measurements described by Shannon. This insight can be seen in Kelly's 1956 paper. [6] While Shannon was interested in entropy primarily with respect to how it affected channel capacity, Kelly suggests that if one associates some utility with an outcome of the stochastic process of message generation then one could compute things like expected value. Note here that message generation is treated as a stochastic process and the idea that individual message might be crafted for some purpose or to have some effect on the recipient is ignored. Again, semantics is irrelevant to the engineering problem in this work.

In Maron's 1960 paper, the notion of utility associated with the outcome of the statistical pro-

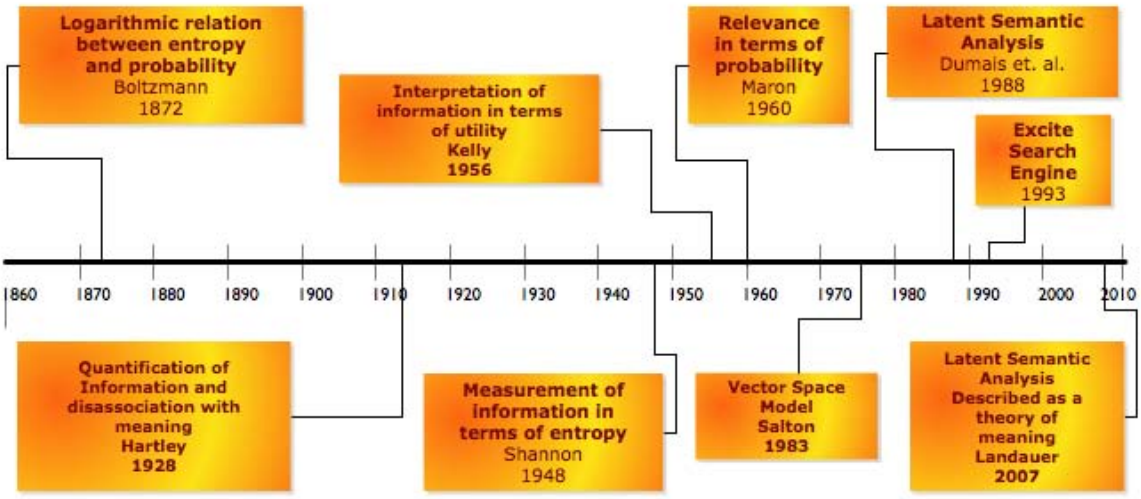

Figure 2.1. Major advances leading to statistical text analysis

cess of message generation is tied down specifically to text analysis in the form of relevance. Maron writes (emphasis mine):

> "The notion of relevance is taken as the key concept in the theory of information retrieval and a comparative concept of relevance is explicated in terms of the theory of probability. The resulting technique called "Probabilistic Indexing," allows a computing machine, given a request for information, to make a statistical inference and derive a number (called the "relevance number") for each document, which is a *measure of the probability that the document will satisfy the given request.* " [9]

So, given some information need, we compute the probability that some document is likely to meet that need and then choose the documents with the highest probability. But all this is accomplished without analyzing document meaning.

Salton and Statistical Text Processing

In the document term matrix method associated with Salton, these previous ideas are consolidated into a process where one can start with documents and end with a set of data structures that lets one query documents and compare them to one another.

Shannon focused on the English language as a whole in his treatment of the statistical properties of text. What he does not deal with is that fact that there is, in fact, not a single distribution of letters and terms in the English language. He was less interested in this because of the need to develop communication systems that could deal with arbitrary messages.

Actual term distributions are skewed with respect to topic areas. So, for example, in the English language as a whole, the term "computer" might occur infrequently. On the other hand, in the proceedings of the American Association of Artificial Intelligence conference (AAAI), the term occurs quite frequently. This topic-based skew in term distribution is what makes text search via statistical analysis possible.

A full discussion of how a document/term matrix and associated algorithms work is beyond the scope of this paper and is sufficiently addressed in other literature. What concerns us here is that although the bag of words methodology represents an advance over Shannon's original work, the advance still does not deal with the semantics of the documents.

Rather than deal with semantics, the document term matrix approach focuses on differentiation. Each unique term in each document in a corpus gets a value that represents how well that term distinguishes the document from other documents. This value is typically the product of two other values, the terms "local" weight in the document and the term's "global" weight in the corpus.

"Global" weights are those that indicate how good a term is as an indexing term. Shannon's entropy measure serves as one of the most popular ways to compute global weights. Terms that have a high entropy (i.e. are spread out evenly across the corpus) are the ones that, in general, are considered poor indexing terms because they fail to distinguish documents from one another. They thus receive a low global weighting. A term that has a low entropy (and thus a high global weighting) is one that occurs infrequently or in an isolated number of the documents. Occurring infrequently in the corpus as a whole is an indication that the term is good at distinguishing the documents in which it occurs.

The "local" weight computed from each document is typically a function of the number of times the term occurs in the document. In an individual document, the terms that occur frequently are considered to characterize the document than terms that occur infrequently.

This methodology works well for keyword searching. But note that nowhere is the meaning of the terms or documents directly addressed. Rather, statistical artifacts are treated as surrogates to let the algorithms compute a kind of expected value of a document against some query. But the ability to differentiate some data set items from one another is very different from the ability to understand the content of the documents themselves. The algorithm is not explicitly finding documents on the topic that the user suggested via keyword. The algorithm is finding documents that contain the keywords and differentially weighting the result based on the keywords that are the most unique.

It should also be noted that these statistical artifacts are rough approximations for differentiating documents from one another. In practice, some terms that have a high entropy across a corpus are quite meaningful in that they are good indicators of the general topics in the corpus. Also,

although theoretically terms with low meaning (stop words like "the" and "and") should wash out of the statistics, they oftentimes don't and it is common to remove them as a step in document processing.

Dumais/Landauer and LSA

Singular value decomposition takes the document term matrix idea one step further, exposing implicit relationships among terms and documents. A full discussion of the details of LSA is beyond the scope of this paper. The reader who wants to understand more should refer to [3]. A good graduate student level treatment of the subject, discussing the whole process from raw text documents to search engine, can be found here [2].

Proponents of LSA, as the name implies, assert that "semantics" are extracted that are latent in the original log-entropy generated matrix. Two primary arguments are made to support this. First, it is known that as children acquire language, their rate of word acquisition accelerates at a rate that does not match the frequency with which they hear the words. In other words, as one acquires language, new words can be mastered faster. Landauer's suggestion [8] is that although the brain certainly is not performing linear algebra, it behaves as if it is accomplishing some kind of dimensionality reduction so that new words can be mastered according to the words that have been learned in the past. If the brain is doing some kind of dimensionality reduction, then doing it in software is psychologically plausible.

More recently, some of the core LSA researchers have made the stronger assertion that LSA can be thought of as a theory of meaning [7]. Their main argument is based on performance. In so far as the performance of a computer system matches that of a person, they argue that LSA is an accurate model of human meaning.

Landauer's argument is clear, concise, and makes sense within its stated scope. That scope, however impressive, is severely limited. Like the term "information" for Shannon, the term "semantics" for LSA advocates has a different definition than the commonly accepted one.

The underlying algorithms only take into account the number of times each term occurs in each document and compares those measurements among all the terms in all documents. As long as that measure stays the same, the resulting analysis stays the same. If, for example, one were to take a corpus and scramble all the words in random order, or place them all in alphabetical order, the analysis won't change. So, for example, the phrase "Sally shot Larry" and the phrase "Larry shot Sally" have the same "semantics" for LSA. Also, for a sufficiently large corpus, the phrases "Bill ate tomatoes" and "Bill ate no tomatoes" have almost exactly the same meaning for LSA.

But when we talk about semantics in general, people often mean the ability to differentiate and interpret these phrases. People can differentiate the phrases because they understand the "meaning" of the system. A computer can't because it only deals with word counts rather than the meaning.

For search engines, this ability is often not terribly relevant. We use the search engine to bring

back documents on a topic and then read the specifics ourselves. But for many of the applications we are targeting today, this difference is critical. We want the algorithms to not only find relevant documents in some broad category, we want algorithms to examine the content of the documents and discover relevant fact that might constitute the "needle in the haystack" or "actionable intelligence." LSA is simply not suited to that problem.

In conclusion, advances beyond Shannon do not address semantics in the sense that is needed for many of today's problems.

This page intentionally left blank.

Chapter 3

How to Re-Introduce Semantics

We now turn our attention to how semantics could be reintroduced at a theoretical level and how that could lead to new algorithms.

What is meant by semantics?

By "semantics," for the purpose of this paper, we mean an effect or an intended effect that the message is supposed to have on the recipient. In the case of "actionable intelligence" or "needle in the haystack," the semantics of the information sought has subtly different goals, but they share one thing in common: the information sought is intended to effect some kind of change on the recipient. In the case of "actionable intelligence," the recipient has one of several different possible actions to choose from. The recipient knows some but not necessarily all of those actions. The new information is intended to alter their state so they can choose among alternatives.

In the case of the needle in the haystack, the recipient has a very large amount of potential information to choose from, most of which is either useless or already known. Some information, however, could be received that would change their state in some useful way.

In order to do this, we have to represent the effect of the message on the recipient. Shannon's theory of communication is concerned solely with the transmission of the message from one point to another. However, by extending it to include the state of the recipient and the receiver, we can account for the effect on the receiver.

New Block Diagram

This transactional definition of semantics can be described by Figure 3.1.

The core idea of this diagram is that two entities, A and B, are sending messages back and forth with one another and these messages may affect one another's state. There are a couple of aspects to this problem that should be pointed out.

In this model, we treat the transmission of symbols as a solved problem. Shannon's diagram

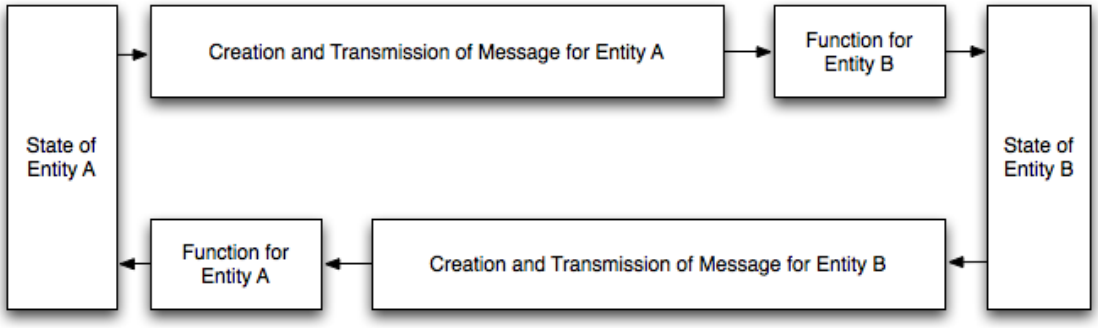

Figure 3.1. Diagram for Transactional Understanding of Communication

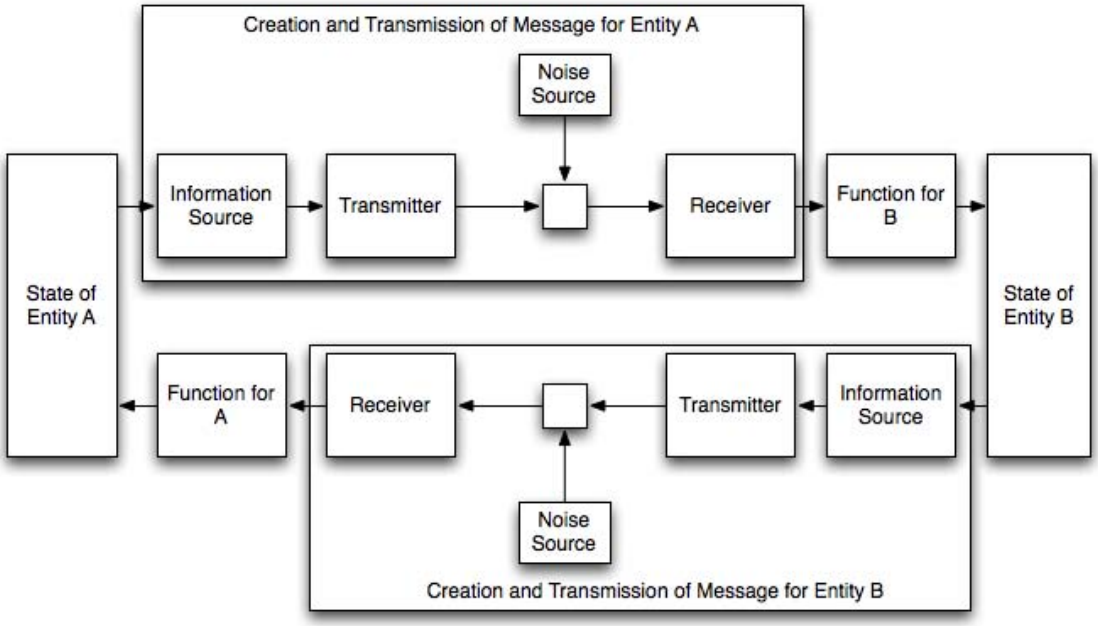

Figure 3.2. The above transactional diagram can be understood as an extension of Shannon's original diagram

itself is still here, but it is bound up in the Creation and Transmission of Message for Entity (A or B). Rather, we focus on the effect that the message will have on the recipient. Figure 3.2 shows how Shannon's model fits in two places, one for messages passing in each direction.

Each entity can send and receive messages. Each entity receives a message by way of a state change function. Each of the two entities has a function that translates the message received into some change of state for the entity. These functions takes the message and the current state of the recipient and modifies the state of the recipient as a result. This means that the same message delivered more than once is not guaranteed to have the same effect each time.

These state change functions are not necessarily the same for each entity. So the same message received by different entities in the same state may still have different effects.

Neither the state nor the functions are necessarily known to the other entity. This uncertainty introduces a number of challenges for estimating the state of the recipient and for crafting messages that reflect real world problems.

As a whole, this model makes a place for context. Rather than focusing on the message itself, the focus shifts to the effect that the message is going to have and thus to the context within which the message is generated and received.

This page intentionally left blank.

Chapter 4

Moving Forward

The goal of this document has been to explore Shannon"s Theory of Communication, specifically his underlying model and to develop and extension to that model that could provide an intellectual foundation to encourage new ways of thinking about the information processing to develop new algorithms and tools. In the previous section, a new model was suggested and discussed. In this section, some potential implications of that model are suggested and some new algorithm work is mentioned that falls within that framework.

Focus on "noise" rather than "signal"

Probably one of the most basic implications of this model is that it would be advantageous to focus on what is generally thought of as the "noise" in today's algorithms. Often, in today's algorithms, messages in data sets can be thought of as being composed of "noise" and "signal." The "signal" is the predictable, repeating part of the data sets that can be detected and extracted. The "noise" is the part of the message that does not correspond to some regularity.

For physical systems that accomplish some overall goal in spite of random variations in its underlying components, this distinction makes sense. But communication is different. In communication, the messages are crafted not to accomplish some single overall emergent thing, but rather to affect some change on the recipient. Redundancy and predictability is included in message for the purpose of overcoming inefficient communication channels. [1] Regularity does not drive the meaning of the message.

The implication is that rather than eliminating "noise" in the analysis process, the noise should be the focus. To put it more succinctly, the noise *is* the signal.

Centrality of user modeling for tool development

We often treat the development of analysis algorithms separately from the engineering of tools that utilize those algorithms and the presentation of the output of those algorithms to end users. So, for example, a new data analysis algorithm may be developed and its success measured according

[1] This goes back to Shannon's theory

to some "ground truth" data. A subsequent step is to create applications and visualizations for displaying the output of the algorithm. The third step is then to marry that analysis capability to an analysis need and try to expose it in a way that an analyst can use.

What often happens in this situation is the proverbial hammer in search of a nail or the proverbial valley of death. The analysis capability works against some ground truth and gives good results, but how it can actually have utility for some individual is unclear.

What this model suggests is that rather than focusing on ground truth performance as a measure of success, algorithm development can start with potential change in state of the user as the goal. This is different from a "user centered" design approach. It isn't only that the end user is part of the environment in which the application is built. The state of the user is an integral part of what the algorithm itself is targeting. The change in user state is the ground truth against which the algorithm is developed.

Focus on the state of the participants

The term *Data Mining* implies that there are some nuggets of truth in a data set that needs to be discovered and extracted. By analogy to actual mining of minerals, the ore would be something that is uniform and can be extracted from the surrounding impurities by various processes.

In the proposed model, the semantics are not contained primarily in the messages like an ore. Rather, the semantics are how the messages change the receipient and in the changes that the sender intends on the recipient. So algorithm explorations using this proposed model should focus on modeling the sender and recipient and how they can be changed as an integral part of data mining.

In the current state of the art, there is much modeling. However what is most common is to treat the model of an individual as just a set of messages (such as a corpus of things written by an individual). Such modeling should also include information about how an individual might change and the kinds of changes they might effect on others.

Inferring the message sender's model of the recipient

In many scenarios where this model might apply, the states of the recipients are partially hidden. Consider a game like chess. Assuming that both players understand the rules to an equal extent, there is some perfectly shared state that is mediated by the game board. However, there is also hidden state in the form of strategy and goals of each player.

Each player also maintains some idea of what the other player is thinking (e.g. my opponent advanced her pawn in an attempt to control the center of the board). That belief constitutes one player's model of the other player. By examining how the player responds to various moves, it may be possible to infer what one player's model of the other player is. Such an examination could

constitute a fruitful line of research.

Inferring the model of the sender

Finally, a study of the model of the sender could also be a fruitful line of research. This is really a superset of the previous point. Formalizing the study of an individual and their state based on a study of the messages produced could also be a fruitful line of study.

This page intentionally left blank.

Chapter 5

Conclusion

The goal of this paper has been threefold. First, we discussed Shannon's theory of communication and subsequent advances leading to modern statistical text analysis, showing that the semantics of messages is ignored. This neglect of semantics makes it difficult for today's algorithms and approaches to address today's problems. Secondly, an alternative model was presented and discussed. Finally, several lines of research were proposed that follow from the proposed model.

This page intentionally left blank.

References

[1] Travis Bauer and David Leake. Using document access sequences to recommend customized information. *IEEE Intelligent Systems Special Issue on Information Customization*, 2002.

[2] Michael W. Berry and Murray Browne. *Understanding search engines : mathematical modeling and text retrieval*. Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1999.

[3] S. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the Society for Information Science*, 41(6):391–407, —1990—. InformationTheory.

[4] F. Ellersick. A conversation with Shannon, Claude. *IEEE Communications Magazine*, 22(5):123–126, —1984—. ISI Document Delivery No.: SQ798 Times Cited: 0 IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC EntityRelationshipExtraction.

[5] R. V. L. Hartley. Transmission of information. *Bell Syst Tech J*, —1928—.

[6] Jr. Kelly, J. A new interpretation of information rate. *Information Theory, IEEE Transactions on*, 2(3):185–189, —1956—. InformationTheory.

[7] T. K. Landauer. LSA as a theory of meaning. In T. K. Landauer, D. McNamara, S. Dennis, and W. Kintsch, editors, *Handbook of Latent Semantic Analysis*, pages 3–34. Lawrence Erlbaum Associates, Mahwah, NJ, —2007—.

[8] Thomas K. Landauer and Susan T. Dumais. A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[9] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *J. ACM*, 7(3):216–244, —1960—. InformationTheory.

[10] E. M. Rogers. Claude Shannon cryptography research during World War II and the mathematical theory of communication. In L. D. Sanson, editor, *IEEE 28th Annual 1994 International Carnahan Conference on Security Technology*, pages 1–5, Albuquerque, Nm, —1994—. I E E E. ISI Document Delivery No.: BC12L Times Cited: 0 EntityRelationshipExtraction.

[11] CE Shannon. A mathematical theory of communication. *Bell Syst Tech J*, 27:379–423,623–656, —1948—. Misc InformationTheory.

[12] W. Weaver. Recent contributions to the mathematical theory of communication. In C. E. Shannon and W. Weaver, editors, *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, —1949—. EntityRelationshipExtraction InformationTheory.

This page intentionally left blank.

Appendix A

Elimination of Time

There is another important aspect of Shannon's 1948 paper that is not explicitly stated, but is has an important influence on his theory and thus on subsequent algorithms. This is the elimination of time. The elimination of time can be seen in two ways.

First, Shannon discusses channel capacity in terms of a rate at which terms can be transmitted. The capacity of a channel is related to the redundancy in the underlying language used to generate the messages. Computing the redundancy in a language in general requires ignoring any individual messages and just looking at the statistical characteristics of the aggregate. For this reason, the timing and order of the information becomes unimportant.

Secondly, Shannon's model is uni-direction. There is a single source and a single destination. By eliminating from the model a potential response to the message, communications is treated as a single transaction. This is not the analysis of a dialog.

Timing and order have played a role in computing trending and other problems that have subsequently been addressed. So time has been utilized, but in a curious way. Rather than dealing with time and order directly, time is converted into some numerical value, such as a time stamp or a sequence number. These static values, now quite void of time, are used as surrogates for time.

One interesting line of research is not to abstract time from the data and simply treat is as a dimension along which measurements are taken, but use time itself as part of the algorithm. For an example where this has been done, see [1].

This page intentionally left blank.

Appendix B

Ontologies

The goal of this paper is to discuss information theory specifically with respect to statistical text analysis. However, ontology work provides an interesting approach that is consistent with Shannon's model but deals with the problem of semantics in a different way than proposed in this paper. So a brief treatment of it seems appropriate.

Ontology research and technology is a popular method for encoding meaning. In ontology work, information is represented as a graph, where vertices represent entities and links represent relationships. Unlike simple graphs, however, ontologies use highly structured and well defined representations for the content of the vertices and links that include additional details about the relationships and entities.

By providing enough information and structure that logical reasoning algorithms can be applied, it is possible to use automated reasoners to process ontologies and infer new facts.

From the perspective of Shannon's theory of communication, ontologies deal with issues of semantics by choosing a detailed, well specified alphabet for the messages. In Shannon's theory, both the originator and the receiver of the messages have to agree on an alphabet. But the nature of the alphabet and the interpretation that either side gives to the alphabet is outside the scope of his paper. Ontologies are essentially special restrictions on the kinds of alphabets that can be chosen such that the elements have an exact meaning to both the originator and the receiver. Effectively, they take Weaver's suggestion that semantics be treated as another kind of alphabet.

Ontologies fail in a number of key areas. First, building an accepted ontology for any non-trivial domain is intractable. Ontology efforts often don't come to completion and rarely capture all of the relevant information needed for their purpose except in highly structured domains. This is the same problem that arises in production/expert systems and is caused by the same problem. The world of human knowledge does not appear to be easily reducible to the kinds of data structures that are easily processed by modern computers.

Secondly, ontologies are not generally reproducible. There is a whole subfield in ontology research trying to address the fact that if any two groups independently try to build an ontology of the same domain, they will inevitably arrive at different ontologies.

These facts have led many people to assert that the term "ontology" is a poor one for the efforts that go under that name. Ontologies not descriptions of the world. Ontologies are context

dependant, including the background of the people building them and the purposes for which they will be applied.

But because the underlying theory does not support the contextual nature of this kind of work, ontologies do not, by necessity, include context, and thus arises the problems described above. A theory that includes context and intent as a critical, base level part of the system could help resolve these problems.

DISTRIBUTION:

1	MS 1235	Travis Bauer, 5635
1	MS 0899	Technical Library, 9536 (electronic)
1	MS 0123	D. Chavez, LDRD Office, 1011

This page intentionally left blank.

Sandia National Laboratories