An Orthonormal Basis for Entailment

Andrew Olney and Zhiqiang Cai

Institute for Intelligent Systems, University of Memphis
365 Innovation Drive, Memphis, TN 38152
aolney@memphis.edu, zcai@memphis.edu

Abstract

Entailment is a logical relationship in which the truth of a proposition implies the truth of another proposition. The ability to detect entailment has applications in IR, QA, and many other areas. This study uses the vector space model to explore the relationship between cohesion and entailment. A Latent Semantic Analysis space is used to measure cohesion between propositions using vector similarity. We present perhaps the first vector space model of entailment. The critical element of the model is the orthonormal basis, which we propose is a geometric construction for inference.

Introduction

Entailment is a logical relationship in which the truth of one or more premises implies the truth of some conclusion. While there are several such possible logical relationships, e.g. the truth of premises guarantees the falsity of a conclusion, or the truth-values of premises and conclusion have no relationship to each other, entailment has a central role in deductive logic. The practical applications for textual entailment are enormous, ranging across question answering, information retrieval, text summarization – any task requiring knowledge representation and inference.

Knowledge representation and inference are two necessary ingredients for entailment. Consider for example the most basic rule of inference, modus ponens, which asserts, "If A implies B and A is true, then B is true." This rule has two kinds of elements, the propositions A and B and the relationship A implies B. Computationally, one may implement such an inference rule using a Post system (Salomaa 1985). In a Post system, sentential propositions are denoted by a single symbol and rules of inference are rewriting rules whereby the premise symbols may be rewritten with a valid conclusion symbol. Many present day systems implementing entailment have a common lineage with Post systems and thus have similar properties.

Unfortunately, problems abound when knowledge and inference are represented in this way. Just as there are an infinite number of sentences in the English language, there are an infinite number of propositions about the world, so creating a complete non-generative knowledge representation is an impossible task. Even with a generative knowledge representation, one must obtain an initial core representation, a non trivial task requiring many thousands of man hours, as witnessed by the Cyc and Wordnet projects (Lenat 1995; Lennat, Miller, and Yokoi 1995; Miller 1995). Finally, symbolization introduces a mapping problem between an inherently fuzzy natural language expression and a precise formal representation, such that some possible interpretations of the natural language statement are invariably lost.

We argue that an ideal knowledge representation for entailment should be generative, use machine learning to bootstrap itself, and retain the inherent fuzziness of natural language. Latent Semantic Analysis (LSA) is a statistical technique for representing world knowledge (Landauer and Dumais 1997; Dumais 1993). LSA is generative, in that any word may be represented by the dimensions of a fixed size vector. Not only does LSA use machine learning, but it also has been shown to closely approximate vocabulary acquisition and usage in children (Landauer and Dumais 1997). Although no work has used LSA to perform entailment per se, LSA has been used to model cohesion and coherence, and the two are closely related.

Foltz, Kintsch, and Landauer (1998) use cohesion to predict comprehension. The relationship between cohesion and comprehension centers on inferences required to maintain coherence. The van Dijk & Kintsch model of comprehension postulates three levels of text representation, the surface code, the text base, and the situation model (van Dijk and Kintsch 1983). Propositions in the text base are derived from the surface code of words, and these propositions are related to each other via semantic coherence and cohesion. Without semantic coherence and cohesion, the reader must draw on her situation model of world knowledge and create bridging inferences that make the propositions coherent. Thus, the Kintsch model predicts that low coherence and cohesion make comprehension more difficult. Foltz, Kintsch, and Landauer (1998) found that LSA makes simple bridging inferences in addition to detecting lexical cohesion. These bridging inferences are a kind of collocational cohesion (Halliday and Hassan 1976) whereby words that co-occur in similar contexts become highly related in the LSA space.

The present study builds upon the idea that LSA has some inferential properties and explores the relationship between cohesion and entailment. Towards this end a LSA space representing general world knowledge was created to measure cohesion and other vector metrics between entailment pairs. Orthonormal bases of LSA vectors are introduced as an important source of metrics for entailment. The study suggests that orthonormal bases are a sufficient condition for entailment using a vector space model.

Section 1 introduces LSA as a vector space model, Section 2 summarizes the use of orthonormal bases with LSA and their relationship to inferencing, Section 3 presents the method used in this study, Section 4 presents the results, Section 5 is the discussion, and Section 6 concludes.

The Vector Space Model

The vector space model is a statistical technique that represents the similarity between collections of words as a cosine between vectors (Manning and Schutze 2002). The process begins by collecting text into a corpus. A matrix is created from the corpus, having one row for each unique word in the corpus and one column for each document or paragraph. The cells of the matrix consist of a simple count of the number of times word *i* appeared in document *j*. Since many words do not appear in any given document, the matrix is often sparse. Weightings are applied to the cells that take into account the frequency of word *i* in document *j* and the frequency of word *i* across all documents, such that distinctive words that appear infrequently are given the most weight. Two collections of words of arbitrary size are compared by creating two vectors. Each word is associated with a row vector in the matrix, and the vector of a collection is simply the sum of all the row vectors of words in that collection. Vectors are compared geometrically by the cosine of the angle between them.

LSA (Landauer and Dumais 1997; Dumais 1993) is an extension of the vector space model that uses singular value decomposition (SVD). SVD is a technique that creates an approximation of the original word by document matrix. After SVD, the original matrix is equal to the product of three matrices, word by singular value, singular value by singular value, and singular value by document. The size of each singular value corresponds to the amount of variance captured by a particular dimension of the matrix. Because the singular values are ordered in decreasing size, it is possible to remove the smaller dimensions and still account for most of the variance. The approximation to the original matrix is optimal, in the least squares sense, for any number of dimensions one would choose. In addition, the removal of smaller dimensions introduces linear dependencies between words that are distinct only in dimensions that account for the least variance. Consequently, two words that were distant in the original space can be near in the compressed space, causing the inductive machine learning and knowledge acquisition effects reported in the literature (Landauer and Dumais 1997).

An Orthonormal Basis

Cohesion can be measured by comparing the cosines of two successive sentences or paragraphs (Foltz, Kintsch, and Landauer 1998). However, cohesion is a crude measure: repetitions of a single sentence will be highly cohesive (cosine of 1) but inadequate since no new information is introduced. A variation of the LSA algorithm using orthonormalized vectors provides two new measures, informativity and relevance, which can detect how much new information is added and how relevant it is in a context (Hu et al. 2003). The essential idea is to represent context by an orthonormalized basis of vectors, one vector for each utterance. The basis is a subspace of the higher dimensional LSA space, in the same way as a plane or line is a subspace of 3D space. The basis is created by projecting each utterance onto the basis of previous utterances using a method known as the Gram-Schmidt process (Anton 2000). Since each vector in the span is orthogonal, the basis represents all linear combinations of what has been previously said. For example, in Figure 1, a new utterance creates a new vector that can be projected to the basis, forming a triangle. The leg of the triangle that lies along the basis indicates the relevance of the recent utterance to the basis; the perpendicular leg indicates new information. Accordingly, a repeated utterance would have complete relevance but zero new information.

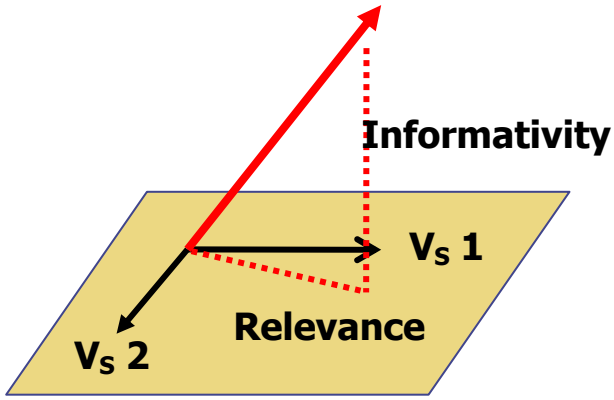

Figure 1. Projecting a new utterance to the basis

Similarly to van Dijk & Kintsch's model of comprehension (van Dijk and Kintsch 1983), dialogue can require inference to maintain coherence. According to Grice's Co-operative Principle, utterances lacking semantic coherence flout the Maxim of Relevance and license an inference (Grice 1975):

S1: Do you want to go out dancing tonight?

S2: I have an exam tomorrow.

The "inference" in the sense of Foltz, Kintsch, and Landauer (1998) is modeled by a fair cosine between these utterances, even though they don't share any of the same words (i.e. the inductive property of LSA). Since the addition of vectors can arbitrarily change the cosine, such inferences can be lost with the addition of more words to a text unit. Using a span, each utterance is kept independent, so inferencing can extend over both the entire set of utterances and the linear combination of any of its subsets. Accordingly, when heavy inferencing is required, one would expect a span-based method to be better able to model semantic coherence and cohesion than a standard vector-based method.

Method

In this section we present the method used in this study. Our first concern was to find a data set for training and testing. A data set has been recently made available via the PASCAL Textual Entailment Challenge, a program which seeks to provide a common data set for comparing applications that perform semantic inference (PASCAL 2004). Below are two example propositions and their corresponding true and false hypothesis:

Eyeing the huge market potential, currently led by Google, Yahoo took over search company Overture Services Inc last year.

Yahoo bought Overture.

Microsoft's rival Sun Microsystems Inc. bought Star Office last month and plans to boost its development as a Web-based device running over the Net on personal computers and Internet appliances.

Microsoft bought Star Office.

We collected all proposition-hypothesis pairs from the public data sets and randomly assigned these to training and testing conditions. After assignment there were 259 pairs in the training set and 308 pairs in the testing set. Amongst all these pairs, approximately half are true entailments and the other half false entailments. None of these pairs were used in the creation of the LSA space, as that could potentially contaminate the results.

The LSA space was created using the Touchstone Applied Science Associates (TASA) corpus used for The Educator's Word Frequency Guide (TASA 2004). TASA constructed the corpus using a “degrees of reading power” scale (DRP) that scores the difficulty of text encountered by students at grades 3, 6, 9, 12, and college. The TASA corpus contains example texts at each of these levels, but more importantly, the DRP scale suggests that the text for a 9th grade student also contain the text for the lower levels. Thus the TASA corpus represents a developmental trajectory in world knowledge, as opposed to WordNet glosses (Miller, 1995) or other textual knowledge representations like the MIT OpenMind CommonSense project (Singh et al., 2002). It is hypothesized in the current study that a developmental corpus will allow better bootstrapping of world knowledge than a non-developmental corpus, an assumption that requires more research.

The present study used the 12th grade version of the TASA corpus. The corpus contains 28,882 documents and 76,132 terms. The corpus was made lower case, and split into paragraphs for LSA. The following example paragraph is typical of this corpus:

a circle of seasons ; livingston, myra cohn ; holiday house ; 1982 26pp. isbn: 0-8234-0452-8 ; drp mean: n/a ; a cycle of poems about the four seasons, illustrated with paintings by leonard everett fisher.

No other processing was performed, e.g. tagging. However, the LSA space was made without counting the standard set of common words, as these are generally thought to only add noise. The space was made with 300 dimensions using log entropy weighting.

Training consisted of calculating various vector space metrics on the proposition-hypothesis pairs and using logistic regression to fit a classifier to the data. Of particular interest are the standard geometric cosine and two new metrics associated with the orthonormal basis. These two metrics are based on the following procedure. First, for each word in the proposition, the associated vector was projected onto the basis, and the component of the vector perpendicular to the basis was normalized and added to the basis. Secondly, the vector representation of the hypothesis was projected into the basis, yielding two elements, a perpendicular component and a parallel component. Given these two elements, the most sensible theoretical metrics are the perpendicular length and cosine between the hypothesis and the parallel component. These reflect the weight of new information in the hypothesis and the similarity of the hypothesis, respectively. These metrics are logical from the standpoint that we expect a hypothesis to be similar to the proposition but at the same time add a little bit of new information. Parameters for these metrics were fitted to a linear model using binary logistic regression, and these models were tested for statistical significance.

Finally, the regression model created in training was evaluated against the unseen testing data. The outcomes of training and testing are reviewed in the next section.

Results

In training we focused on both cosine and the two basis metrics given above. Using a χ^2 test for significance, we found that the regression model using cosine is not a significant predictor of entailment at $p < .05$. However,

the regression model using both basis metrics is a significant predictor of entailment at $p < .05$. The error confusion matrix for training is given below:

	Observed	
Predicted	True	False
True	95	73
False	40	51

Table 1. Error confusion matrix for training

While precision, recall, and F-measure (Manning and Schutze) could be calculated for this error-confusion matrix, these measures are misleading in this context. Consider that a model which always predicts TRUE will have an F-measure of .67 out of 1.0, because the TRUE and FALSE cases have even odds. In this situation, d-prime and the associated measures of hit rate and false alarm rate are more informative (Swets, Tanner, and Birdsall 1961). For the regression model using the orthonormal basis, these are .70, .59, and .31 respectively. With respect to a baseline of chance, the d-prime indicates that there are .31 normal standard deviates between the mean of the non-entailment distribution and the mean of the entailment distribution. The Receiver Operating Characteristic Curve in Figure 2 illustrates the relationship between hit rate and false alarm rate in the model.

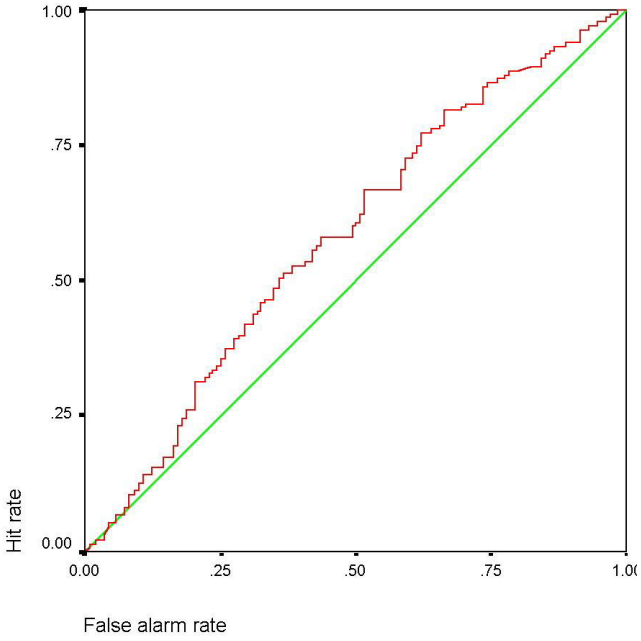

Figure 2. Receiver Operating Characteristic Curve

Binomial logistic regression produced a model which predicts TRUE when the following value is greater than .5:

$$\frac{e^{-.275-1.429*perpendicularLength+.667*parallelCo\sin e}}{1+e^{-.275-1.429*perpendicularLength+.667*parallelCo\sin e}}$$

The parameters of this equation are a constant term, -.275 and two coefficients for the basis metrics, -1.429 and .667. The sign of these last two coefficients illuminates their role in the model. The negative sign on the coefficient for perpendicular length indicates that perpendicular length is inversely correlated with a hypothesis being a valid entailment. Therefore too much new information in a hypothesis contraindicates a valid entailment. On the other hand, the positive sign on the parallel cosine indicates that a hypothesis should have elements in common with its proposition. Beyond the sign, the exponentials of these values indicate how the two basis metrics are weighted. For example, when perpendicular length is raised by one unit, a FALSE entailment is .24 more times likely. When parallel cosine is raised by one unit, a TRUE entailment is 1.95 more times likely.

Running the trained model over the testing data showed that the correct classification was significant, $p < .05$, using a χ^2 test for significance. Performance was surprisingly consistent across both training and testing. The testing hit rate, false alarm rate, and d-prime are .72, .60, and .30, compared to the training values of .70, .59, and .31. This is strong evidence that the model has generalized well to the underlying pattern in the data. The error confusion matrix for the test data is given in Table 2.

	Observed	
Predicted	True	False
True	106	97
False	42	63

Table 2. Error confusion matrix for testing

Discussion

Our intuition that the basis method is performing a kind of inference is strongly supported by these results. In the general case for this data set, inference is required to determine whether an entailment relation exists. For if simple substitution were enough, the cosine metric would be a significant predictor of entailment, and we have found that cosine is not a significant predictor. Thus we have shown, perhaps for the first time, that it is possible to use a vector space model by itself to detect entailments. Moreover, we have shown that using the basis method, a general knowledge source like TASA can be generative enough to make inferences over foreign material.

Not only do the results demonstrate that inference is required to solve this problem, but also the manner in which inference is supplied is consistent with the geometric intuition behind the basis model. Recall that

since a basis is a subspace of a vector space, an infinite number of statements can be represented inside the basis. The propositions generated by the basis are a clear analogue to implicatures generated by inference. According to intuition, an entailment should be largely contained within the subspace generated by the basis; otherwise it would not be inferable. Likewise in our basis model, the greater the cosine between the basis and the hypothesis vector, the more likely the hypothesis is to be an entailment (1.95 more times likely for every unit increase of cosine). Symmetrically, the greater the weight of information outside the basis, as represented by the length of the hypothesis component perpendicular to the basis, the less likely the hypothesis is to be an entailment (.24 times less likely for every unit increase in perpendicular length).

Equal training and testing performance suggests that the model is generalizing well with respect to what it can represent. The basis is a intuitive model for inference, and it is possible that low performance speaks more to the kinds of information that are currently missing from the vector space model than from an inherent weakness in the basis method. The orthonormal basis method does quite well, especially considering that it stands on the shoulders of LSA, which only makes use of collocational information in the corpus.

Entailment is a notoriously difficult problem, and there are a number of factors working against the current method. One is that the training and testing set contain many terms that are not found in the LSA space. This is because these data sets are drawn largely from current events and contain many proper nouns, e.g. "Aristide" that are not likely to be found in TASA. In addition, the entailment data set makes use of mathematical notation, relying on a system's ability to reason over quantity. This is not possible with LSA – a mathematical expression or quantity is simply substitutable for another quantity or word. The substitutability issue comes up also with antonyms, synonyms, and hyper/hyponyms, all of which LSA treats the same (Foltz, Kintsch, and Landauer 1998). Consider the following entailment pair taken from the data set:

> Crude oil for April delivery traded at $37.80 a barrel, down 28 cents
>
> Crude oil prices rose to $37.80 per barrel

Clearly "down" and "rose" have an antonym flavor to their relationship, but the LSA cosine between these two words is high because they are substitutable in the same contexts. Finally, word order is a problem in LSA. Because addition is commutative, i.e. 1+2 = 2+1, word order is lost when word vectors are added to create document vectors. Therefore "John likes Mary" and "Mary likes John", which do not have an entailment relationship between them, would have a cosine of 1. All of these factors are challenges for future research.

It is important to recognize that the method outlined in this paper specifically addresses the verification of inferences rather than their generation. However, within the given framework, it is possible to view the orthonormal basis as an approximation of all of the valid inferences obtainable from a statement. Clearly this approximation over-generates, because it pays no attention to word order, negation, or the other phenomena listed above. While this perspective is consistent with the current approach, the present study cannot confirm it.

Conclusion

The vector space model uses machine learning, is generative, and maintains the fuzziness of natural language. As such, the vector space model is a natural candidate to consider for knowledge representation. This study has shown that one can capitalize on vector space structure to create the capability for inference and entailment. The orthonormal basis method is a clean extension to the vector space model with a clear analogue to inference: the basis represents all linear combinations of the words in the hypothesis and as such should largely contain the hypothesis. This study is perhaps the first to demonstrate entailment using the vector space model. New enhancements and extensions await discovery.

Acknowledgements

This research on was supported by the National Science Foundation and the Department of Defense Multidisciplinary University Research Initiative administered by the Office of Naval Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DoD, ONR, or NSF.

References

Anton, H. 2000. *Elementary linear algebra*. 8th edition. New York: John Wiley.

van Dijk, T. A., and Kintsch, W. 1983. *Strategies of Discourse Comprehension*. New York: Academic Press.

Dumais, S. 1993. LSI Meets TREC: A Status Report. In Proceedings of the First Text Retrieval Conference (TREC1), 137-152. NIST Special Publication 500-207.

Foltz, P.W.; Kintsch, W.; and Landauer, TK. 1998. The measurement of textual Cohesion with Latent Semantic Analysis. *Discourse Processes,* 25: 285-307.

Garson, D. Logistic Regression. Accessed on April 18th, 2003. http://www2.chass.ncsu.edu/garson/pa765/logistic

Grice, H.P. 1975. Logic and conversation. In Cole, P. and Morgan, J. (eds) *Syntax and Semantics* 3: 41-58. New York: Academic.

Halliday, M. A. and Hassan, R. A.. 1976. *Cohesion in English*. London: Longman.

Hu, X.; Cai, Z.; Louwerse, M.; Olney, A.; Penumatsa, P.; and Graesser, A. 2003. An improved LSA algorithm to evaluate contributions in student dialogue. In Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03), 1489-1491.

Landauer, T. and Dumais, S. 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104: 211-240.

Lennat, D. 1995. CYC: a large scale investment in knowledge infrastructure. *Communications of the ACM* 38(11): 33-38.

Lennat, D.; Miller, G.; and Yokoi, T. 1995. CYC, Wordnet, and EDR: critiques and responses. *Communications of the ACM* 38(11): 45-48.

Manning, C. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press.

Miller, G. 1995. Wordnet: a lexical database for English. *Communications of the ACM* 38(11): 39-41.

PASCAL. 2004. Recognising Textual Entailment Challenge. Accessed on October 4th, 2004. http://www.pascal-network.org/Challenges/RTE/

Salomaa, A. 1985. *Computation and Automata*. In Rota, G. C. (ed) The Encyclopedia of Mathematics and its Applications, Volume 25. Cambridge: Cambridge University Press.

Singh, P.; Lin, T.; Mueller E.; Lim, G.; Perkins, T.; and Zhu, W. 2002. Open Mind Common Sense: Knowledge acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, Lecture Notes in Computer Science*. Heidelberg: Springer-Verlag.

Swets J.; Tanner W.; and Birdsall T. 1961. Decision processes in perception. *Psychological Review*, 68(5): 301-340.

TASA. 2004. Touchstone Applied Science Associates Corpus. Accessed on October 20th, 2004 http://lsa.colorado.edu/spaces.html