Independence Assumptions Considered Harmful

Alexander Franz
Sony Computer Science Laboratory & D21 Laboratory
Sony Corporation
6-7-35 Kitashinagawa
Shinagawa-ku, Tokyo 141, Japan
amf@csl.sony.co.jp

Abstract

Many current approaches to statistical language modeling rely on independence assumptions between the different explanatory variables. This results in models which are computationally simple, but which only model the main effects of the explanatory variables on the response variable. This paper presents an argument in favor of a statistical approach that also models the interactions between the explanatory variables. The argument rests on empirical evidence from two series of experiments concerning automatic ambiguity resolution.

1 Introduction

In this paper, we present an empirical argument in favor of a certain approach to statistical natural language modeling: we advocate statistical natural language models that account for the interactions between the explanatory statistical variables, rather than relying on independence assumptions. Such models are able to perform prediction on the basis of estimated probability distributions that are properly conditioned on the combinations of the individual values of the explanatory variables.

After describing one type of statistical model that is particularly well-suited to modeling natural language data, called a loglinear model, we present empirical evidence from a series of experiments on different ambiguity resolution tasks that show that the performance of the loglinear models outranks the performance of other models described in the literature that assume independence between the explanatory variables.

2 Statistical Language Modeling

By "statistical language model", we refer to a mathematical object that "imitates the properties" of some aspects of natural language, and in turn makes predictions that are useful from a scientific or engineering point of view. Much recent work in this framework has used written and spoken natural language data to estimate parameters for statistical models that were characterized by serious limitations: models were either limited to a single explanatory variable or, if more than one explanatory variable was considered, the variables were assumed to be independent. In this section, we describe a method for statistical language modeling that transcends these limitations.

2.1 Categorical Data Analysis

Categorical data analysis is the area of statistics that addresses *categorical* statistical variable: variables whose values are one of a set of categories. An example of such a linguistic variable is PART-OF-SPEECH, whose possible values might include *noun, verb, determiner, preposition*, etc.

We distinguish between a set of explanatory variables, and one response variable. A statistical model can be used to perform prediction in the following manner: Given the values of the explanatory variables, what is the probability distribution for the response variable, i.e., what are the probabilities for the different possible values of the response variable?

2.2 The Contingency Table

The basic tool used in categorical data analysis is the contingency table (sometimes called the "cross-classified table of counts"). A contingency table is a matrix with one dimension for each variable, including the response variable. Each cell in the contingency table records the frequency of data with the appropriate characteristics.

Since each cell concerns a specific combination of features, this provides a way to estimate probabilities of specific feature combinations from the observed frequencies, as the cell counts can easily be converted to probabilities. Prediction is achieved by determining the value of the response variable given the values of the explanatory variables.

2.3 The Loglinear Model

A loglinear model is a statistical model of the effect of a set of categorical variables and their combinations on the cell counts in a contingency table. It can be used to address the problem of sparse data, since it can act as a "smoothing device, used to obtain cell estimates for every cell in a sparse array, even if the observed count is zero" (Bishop, Fienberg, and Holland, 1975).

Marginal totals (sums for all values of some variables) of the observed counts are used to estimate the parameters of the loglinear model; the model in turn delivers estimated expected cell counts, which are smoother than the original cell counts.

The mathematical form of a loglinear model is as follows. Let m_{ijk} be the expected cell count for cell $(i, j, k, \ldots)$ in the contingency table. The general form of a loglinear model is as follows:

$$\log m_{ijk\ldots} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + \ldots \quad (1)$$

In this formula, u denotes the mean of the logarithms of all the expected counts, $u + u_{1(i)}$ denotes the mean of the logarithms of the expected counts with value i of the first variable, $u + u_{2(j)}$ denotes the mean of the logarithms of the expected counts with value j of the second variable, $u + u_{12(ij)}$ denotes the mean of the logarithms of the expected counts with value i of the first veriable and value j of the second variable, and so on.

Thus, the term $u_{1(i)}$ denotes the deviation of the mean of the expected cell counts with value i of the first variable from the grand mean u. Similarly, the term $u_{12(ij)}$ denotes the deviation of the mean of the expected cell counts with value i of the first variable and value j of the second variable from the grand mean u. In other words, $u_{12(ij)}$ represents the *combined effect* of the values i and j for the first and second variables on the logarithms of the expected cell counts.

In this way, a loglinear model provides a way to estimate expected cell counts that depend not only on the main effects of the variables, but also on the interactions between variables. This is achieved by adding "interaction terms" such as $u_{12(ij)}$ to the model. For further details, see (Fienberg, 1980).

2.4 The Iterative Estimation Procedure

For some loglinear models, it is possible to obtain closed forms for the expected cell counts. For more complicated models, the *iterative proportional fitting* algorithm for hierarchical loglinear models (Deming and Stephan, 1940) can be used. Briefly, this procedure works as follows.

Let the values for the expected cell counts that are estimated by the model be represented by the symbol $\hat{m}_{ijk\ldots}$. The interaction terms in the loglinear models represent constraints on the estimated expected marginal totals. Each of these marginal constraints translates into an adjustment scaling factor for the cell entries. The iterative procedure has the following steps:

1. Start with initial estimates for the estimated expected cell counts. For example, set all $\hat{m}_{ijkl} = 1.0$.
2. Adjust each cell entry by multiplying it by the scaling factors. This moves the cell entries towards satisfaction of the marginal constraints specified by the model.
3. Iterate through the adjustment steps until the maximum difference ϵ between the marginal totals observed in the sample and the estimated marginal totals reaches a certain minimum threshold, e.g. $\epsilon = 0.1$.

After each cycle, the estimates satisfy the constraints specified in the model, and the estimated expected marginal totals come closer to matching the observed totals. Thus, the process converges. This results in Maximum Likelihood estimates for both multinomial and independent Poisson sampling schemes (Agresti, 1990).

2.5 Modeling Interactions

For natural language classification and prediction tasks, the aim is to estimate a conditional probability distribution $P(H|E)$ over the possible values of the hypothesis H, where the evidence E consists of a number of linguistic features $e_1, e_2, \ldots$. Much of the previous work in this area assumes independence between the linguistic features:

$$P(H|e_i, e_j, \ldots) \approx P(H|e_i) \times P(H|e_j) \times \ldots \quad (2)$$

For example, a model to predict Part-of-Speech of a word on the basis of its morphological affix and its capitalization might assume independence between the two explanatory variables as follows:

$$P(\mathsf{POS}|\mathsf{AFFIX}, \mathsf{CAPITALIZATION}) \approx \quad (3)$$
$$\approx P(\mathsf{POS}|\mathsf{AFFIX}) \times P(\mathsf{POS}|\mathsf{CAPITALIZATION})$$

This results in a considerable computational simplification of the model but, as we shall see below, leads to a considerable loss of information and concomitant decrease in prediction accuracy. With a loglinear model, on the other hand, such independence assumptions are not necessary. The loglinear model provides a posterior distribution that is properly conditioned on the evidence, and maximizing the conditional probability $P(H|E)$ leads to minimum error rate classification (Duda and Hart, 1973).

3 Predicting Part-of-Speech

We will now turn to the empirical evidence supporting the argument against independence assumptions. In this section, we will compare two models for predicting the Part-of-Speech of an unknown word: A simple model that treats the various explanatory variables as independent, and a model using loglinear smoothing of a contingency table that takes into account the interactions between the explanatory variables.

3.1 Constructing the Model

The model was constructed in the following way. First, features that could be used to guess the POS of a word were determined by examining the training portion of a text corpus. The initial set of features consisted of the following:

- INCLUDES-NUMBER. Does the word include a number?
- CAPITALIZED. Is the word in sentence-initial position and capitalized, in any other position and capitalized, or in lower case?
- INCLUDES-PERIOD. Does the word include a period?
- INCLUDES-COMMA. Does the word include a comma?
- FINAL-PERIOD. Is the last character of the word a period?
- INCLUDES-HYPHEN. Does the word include a hyphen?
- ALL-UPPER-CASE. Is the word in all upper case?
- SHORT. Is the length of the word three characters or less?
- INFLECTION. Does the word carry one of the English inflectional suffixes?
- PREFIX. Does the word carry one of a list of frequently occurring prefixes?
- SUFFIX. Does the word carry one of a list of frequently occurring suffixes?

Next, exploratory data analysis was performed in order to determine relevant features and their values, and to approximate which features interact. Each word of the training data was then turned into a feature vector, and the feature vectors were cross-classified in a contingency table. The contingency table was smoothed using a loglinear models.

3.2 Data

Training and evaluation data was obtained from the Penn Treebank Brown corpus (Marcus, Santorini, and Marcinkiewicz, 1993). The characteristics of "rare" words that might show up as unknown words differ from the characteristics of words in general, so a two-step procedure was employed a first time to obtain a set of "rare" words as training data, and again a second time to obtain a separate set of "rare" words as evaluation data. There were 17,000 words in the training data, and 21,000 words in the evaluation data. Ambiguity resolution accuracy was evaluated for the "overall accuracy" (Percentage that the most likely POS tag is correct), and "cutoff factor accuracy" (accuracy of the answer set consisting of all POS tags whose probability lies within a factor F of the most likely POS (de Marcken, 1990)).

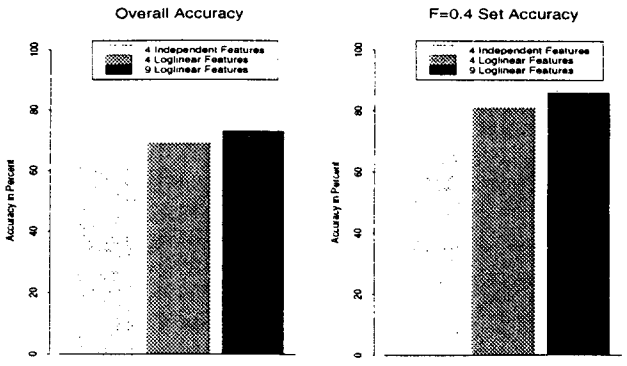

Figure 1: Performance of Different Models

3.3 Accuracy Results

(Weischedel et al., 1993) describe a model for unknown words that uses four features, but treats the features as independent. We reimplemented this model by using four features: POS, INFLECTION, CAPITALIZED, and HYPHENATED. In Figures 1 2, the results for this model are labeled **4 Independent Features**. For comparison, we created a loglinear model with the same four features; the results for this model are labeled **4 Loglinear Features**.

The highest accuracy was obtained by the loglinear model that includes all two-way interactions and consists of two contingency tables with the following features: POS, ALL-UPPER-CASE, HYPHENATED, INCLUDES-NUMBER, CAPITALIZED, INFLECTION, SHORT, PREFIX, and SUFFIX. The results for this model are labeled **9 Loglinear Features**. The parameters for all three unknown word models were estimated from the training data, and the models were evaluated on the evaluation data.

The accuracy of the different models in assigning the most likely POSs to words is summarized in Figure 1. In the left diagram, the two barcharts show two different accuracy measures: Percent correct (**Overall Accuracy**), and percent correct within the F=0.4 cutoff factor answer set (**F=0.4 Set Accuracy**). In both cases, the loglinear model with four features obtains higher accuracy than the method that assumes independence between the same four features. The loglinear model with nine

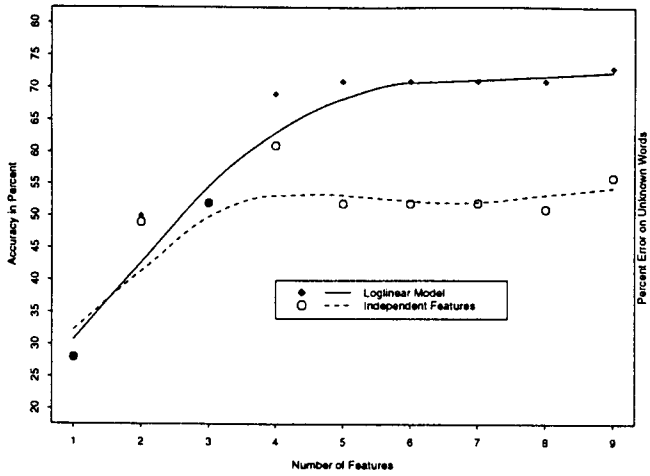

Figure 2: Effect of Number of Features on Accuracy

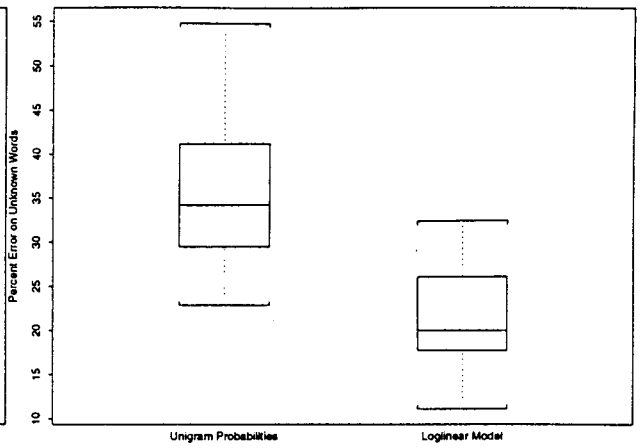

Figure 3: Error Rate on Unknown Words

features further improves this score.

3.4 Effect of Number of Features on Accuracy

The performance of the loglinear model can be improved by adding more features, but this is not possible with the simpler model that assumes independence between the features. Figure 2 shows the performance of the two types of models with feature sets that ranged from a single feature to nine features.

As the diagram shows, the accuracies for both methods rise with the first few features, but then the two methods show a clear divergence. The accuracy of the simpler method levels off around at around 50–55%, while the loglinear model reaches an accuracy of 70–75%. This shows that the loglinear model is able to tolerate redundant features and use information from more features than the simpler method, and therefore achieves better results at ambiguity resolution.

3.5 Adding Context to the Model

Next, we added of a stochastic POS tagger (Charniak et al., 1993) to provide a model of context. A stochastic POS tagger assigns POS labels to words in a sentence by using two parameters:

- **Lexical Probabilities:** $P(w|t)$ — the probability of observing word w given that the tag t occurred.
- **Contextual Probabilities:** $P(t_i|t_{i-1}, t_{i-2})$ — the probability of observing tag t_i given that the two previous tags t_{i-1}, t_{i-2} occurred.

The tagger maximizes the probability of the tag sequence $T = t_1, t_2, \ldots, t_n$ given the word sequence $W = w_1, w_2, \ldots, w_n$, which is approximated as follows:

$$P(T|W) \approx \prod_{i=1}^{n} P(w_i|t_i)P(t_i|t_{i-1}, t_{i-2}) \qquad (4)$$

The accuracy of the combination of the loglinear model for local features and the stochastic POS tagger for contextual features was evaluated empirically by comparing three methods of handling unknown words:

- **Unigram:** Using the prior probability distribution $P(t)$ of the POS tags for rare words.
- **Probabilistic UWM:** Using the probabilistic model that assumes independence between the features.
- **Classifier UWM:** Using the loglinear model for unknown words.

Separate sets of training and evaluation data for the tagger were obtained from from the Penn Treebank Wall Street corpus. Evaluation of the combined system was performed on different configurations of the POS tagger on 30–40 different samples containing 4,000 words each.

Since the tagger displays considerable variance in its accuracy in assigning POS to unknown words in context, we use boxplots to display the results. Figure 3 compares the tagging error rate on unknown words for the unigram method (left) and the loglinear method with nine features (labeled **statistical classifier**) at right. This shows that the loglinear model significantly improves the Part-of-Speech tagging accuracy of a stochastic tagger on unknown words. The median error rate is lowered considerably, and samples with error rates over 32% are eliminated entirely.

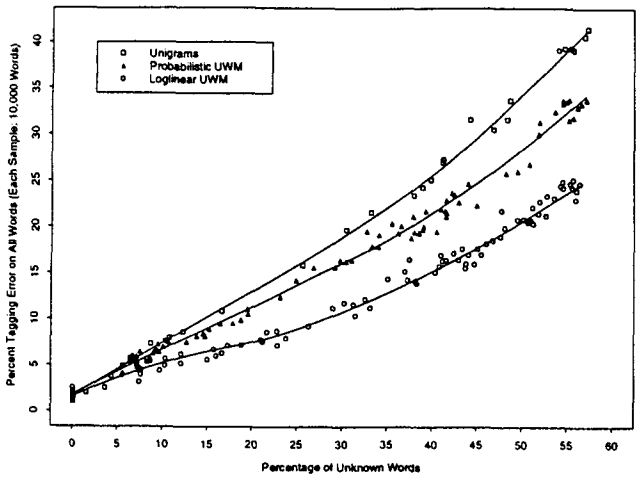

Figure 4: Effect of Proportion of Unknown Words on Overall Tagging Error Rate

3.6 Effect of Proportion of Unknown Words

Since most of the lexical ambiguity resolution power of stochastic POS tagging comes from the lexical probabilities, unknown words represent a significant source of error. Therefore, we investigated the effect of different types of models for unknown words on the error rate for tagging text with different proportions of unknown words.

Samples of text that contained different proportions of unknown words were tagged using the three different methods for handling unknown words described above. The overall tagging error rate increases significantly as the proportion of new words increases. Figure 4 shows a graph of overall tagging accuracy versus percentage of unknown words in the text. The graph compares the three different methods of handling unknown words. The diagram shows that the loglinear model leads to better overall tagging performance than the simpler methods, with a clear separation of all samples whose proportion of new words is above approximately 10%.

4 Predicting PP Attachment

In the second series of experiments, we compare the performance of different statistical models on the task of predicting Prepositional Phrase (PP) attachment.

4.1 Features for PP Attachment

First, an initial set of linguistic features that could be useful for predicting PP attachment was determined. The initial set included the following features:

- PREPOSITION. Possible values of this feature include one of the more frequent prepositions in the training set, or the value *other-prep*.
- VERB-LEVEL. Lexical association strength between the verb and the preposition.
- NOUN-LEVEL. Lexical association strength between the noun and the preposition.
- NOUN-TAG. Part-of-Speech of the nominal attachment site. This is included to account for correlations between attachment and syntactic category of the nominal attachment site, such as "PPs disfavor attachment to proper nouns."
- NOUN-DEFINITENESS. Does the nominal attachment site include a definite determiner? This feature is included to account for a possible correlation between PP attachment to the nominal site and definiteness, which was derived by (Hirst, 1986) from the principle of presupposition minimization of (Crain and Steedman, 1985).
- PP-OBJECT-TAG. Part-of-speech of the object of the PP. Certain types of PP objects favor attachment to the verbal or nominal site. For example, temporal PPs, such as *"in 1959"*, where the prepositional object is tagged CD (cardinal), favor attachment to the VP, because the VP is more likely to have a temporal dimension.

The association strengths for VERB-LEVEL and NOUN-LEVEL were measured using the Mutual Information between the noun or verb, and the preposition.[1] The probabilities were derived as Maximum Likelihood estimates from all PP cases in the training data. The Mutual Information values were ordered by rank. Then, the association strengths were categorized into eight levels (A-H), depending on percentile in the ranked Mutual Information values.

4.2 Experimental Data and Evaluation

Training and evaluation data was prepared from the Penn treebank. All 1.1 million words of parsed text in the Brown Corpus, and 2.6 million words of parsed WSJ articles, were used. All instances of PPs that are attached to VPs and NPs were extracted. This resulted in 82,000 PP cases from the Brown Corpus, and 89,000 PP cases from the WSJ articles. Verbs and nouns were lemmatized to their root forms if the root forms were attested in the corpus. If the root form did not occur in the corpus, then the inflected form was used.

All the PP cases from the Brown Corpus, and 50,000 of the WSJ cases, were reserved as training data. The remaining 39,00 WSJ PP cases formed the evaluation pool. In each experiment, performance

[1] Mutual Information provides an estimate of the magnitude of the ratio between the joint probability P(verb/noun,preposition), and the joint probability assuming independence P(verb/noun)P(preposition) - see (Church and Hanks, 1990).

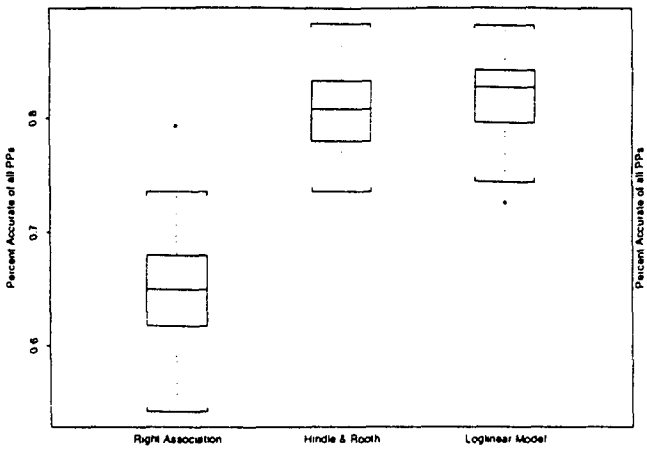

Figure 5: Results for Two Attachment Sites

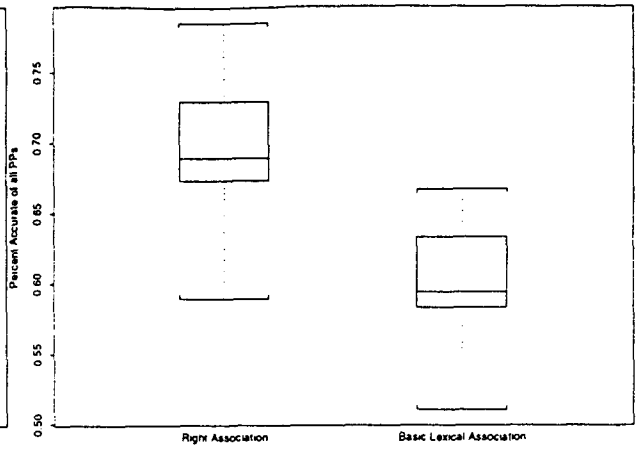

Figure 6: Three Attachment Sites: Right Association and Lexical Association

was evaluated on a series of 25 random samples of 100 PP cases from the evaluation pool, in order to provide a characterization of the error variance.

4.3 Experimental Results: Two Attachments Sites

Previous work on automatic PP attachment disambiguation has only considered the pattern of a verb phrase containing an object, and a final PP. This leads to two possible attachment sites, the verb and the object of the verb. The pattern is usually further simplified by considering only the heads of the possible attachment sites, corresponding to the sequence "Verb Noun_1 Preposition Noun_2".

The first set of experiments concerns this pattern. There are 53,000 such cases in the training data, and 10,000 such cases in the evaluation pool. A number of methods were evaluated on this pattern according to the 25-sample scheme described above. The results are shown in Figure 5.

4.3.1 Baseline: Right Association

Prepositional phrases exhibit a tendency to attach to the most recent possible attachment site; this is referred to as the principle of "Right Association". For the "V NP PP" pattern, this means preferring attachment to the noun phrase. On the evaluation samples, a median of 65% of the PP cases were attached to the noun.

4.3.2 Results of Lexical Association

(Hindle and Rooth, 1993) described a method for obtaining estimates of lexical association strengths between nouns or verbs and prepositions, and then using lexical association strength to predict PP attachment. In our reimplementation of this method, the probabilities were estimated from all the PP cases in the training set. Since our training data are bracketed, it was possible to estimate the lexical associations with much less noise than Hindle & Rooth, who were working with unparsed text. The median accuracy for our reimplementation of Hindle & Rooth's method was 81%. This is labeled "Hindle & Rooth" in Figure 5.

4.3.3 Results of the Loglinear Model

The loglinear model for this task used the features PREPOSITION, VERB-LEVEL, NOUN-LEVEL, and NOUN-DEFINITENESS, and it included all second-order interaction terms. This model achieved a median accuracy of 82%.

Hindle & Rooth's lexical association strategy only uses one feature (lexical association) to predict PP attachment, but, as the boxplot shows, the results from the loglinear model for the "V NP PP" pattern do not show any significant improvement.

4.4 Experimental Results: Three Attachment Sites

As suggested by (Gibson and Pearlmutter, 1994), PP attachment for the "Verb NP PP" pattern is relatively easy to predict because the two possible attachment sites differ in syntactic category, and therefore have very different kinds of lexical preferences. For example, most PPs with *of* attach to nouns, and most PPs with *to* and *by* attach to verbs. In actual texts, there are often more than two possible attachment sites for a PP. Thus, a second, more realistic series of experiments was performed that investigated different PP attachment strategies for the pattern "Verb Noun_1 Noun_2 Preposition Noun_3" that includes more than two possible attachment sites that are not syntactically heterogeneous. There were 28,000 such cases in the training data, and 8000 cases in the evaluation pool.

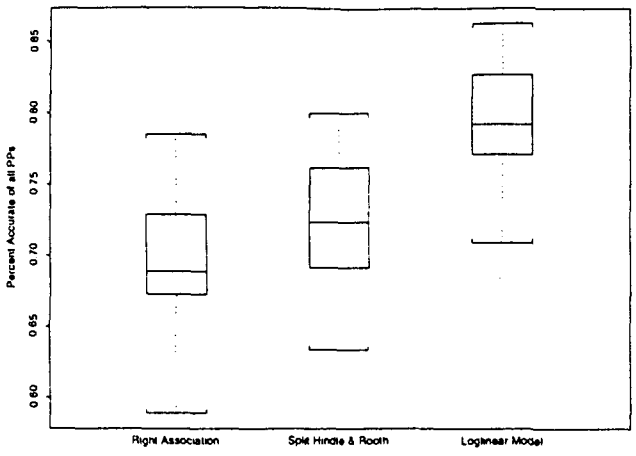

Figure 7: Summary of Results for Three Attachment Sites

4.4.1 Baseline: Right Association

As in the first set of experiments, a number of methods were evaluated on the three attachment site pattern with 25 samples of 100 random PP cases. The results are shown in Figures 6–7. The baseline is again provided by attachment according to the principle of "Right Attachment" to the most recent possible site, i.e. attachment to Noun_2. A median of 69% of the PP cases were attached to Noun_2.

4.4.2 Results of Lexical Association

Next, the lexical association method was evaluated on this pattern. First, the method described by Hindle & Rooth was reimplemented by using the lexical association strengths estimated from all PP cases. The results for this strategy are labeled "Basic Lexical Association" in Figure 6. This method only achieved a median accuracy of 59%, which is worse than always choosing the rightmost attachment site. These results suggest that Hindle & Rooth's scoring function worked well in the "Verb Noun_1 Preposition Noun_2" case not only because it was an accurate estimator of lexical associations between individual verbs/nouns and prepositions which determine PP attachment, but also because it accurately predicted the general verb-noun skew of prepositions.

4.4.3 Results of Enhanced Lexical Association

It seems natural that this pattern calls for a combination of a structural feature with lexical association strength. To implement this, we modified Hindle & Rooth's method to estimate attachments to the verb, first noun, and second noun separately. This resulted in estimates that combine the structural feature directly with the lexical association strength. The modified method performed better than the original lexical association scoring function, but it still only obtained a median accuracy of 72%. This is labeled "Split Hindle & Rooth" in Figure 7.

4.4.4 Results of Loglinear Model

To create a model that combines various structural and lexical features without independence assumptions, we implemented a loglinear model that includes the variables VERB-LEVEL, FIRST-NOUN-LEVEL, and SECOND-NOUN-LEVEL.[2] The loglinear model also includes the variables PREPOSITION and PP-OBJECT-TAG. It was smoothed with a loglinear model that includes all second-order interactions.

This method obtained a median accuracy of 79%; this is labeled "Loglinear Model" in Figure 7. As the boxplot shows, it performs significantly better than the methods that only use estimates of lexical association. Compared with the "Split Hindle & Rooth" method, the samples are a little less spread out, and there is no overlap at all between the central 50% of the samples from the two methods.

4.5 Discussion

The simpler "V NP PP" pattern with two syntactically different attachment sites yielded a null result: The loglinear method did not perform significantly better than the lexical association method. This could mean that the results of the lexical association method can not be improved by adding other features, but it is also possible that the features that could result in improved accuracy were not identified.

The lexical association strategy does not perform well on the more difficult pattern with three possible attachment sites. The loglinear model, on the other hand, predicts attachment with significantly higher accuracy, achieving a clear separation of the central 50% of the evaluation samples.

5 Conclusions

We have contrasted two types of statistical language models: A model that derives a probability distribution over the response variable that is properly conditioned on the combination of the explanatory variable, and a simpler model that treats the explanatory variables as independent, and therefore models the response variable simply as the addition of the individual main effects of the explanatory variables.

[2] These features use the same Mutual Information-based measure of lexical association as the previous loglinear model for two possible attachment sites, which were estimated from all nominal and verbal PP attachments in the corpus. The features FIRST-NOUN-LEVEL and SECOND-NOUN-LEVEL use the same estimates; in other words, in contrast to the "split Lexical Association" method, they were not estimated separately for the two different nominal attachment sites.

The experimental results show that, with the same feature set, modeling feature interactions yields better performance: such models achieves higher accuracy, and its accuracy can be raised with additional features. It is interesting to note that modeling variable interactions yields a higher performance gain than including additional explanatory variables.

While these results do not *prove* that modeling feature interactions is necessary, we believe that they provide a strong indication. This suggests a number of avenues for further research.

First, we could attempt to improve the specific models that were presented by incorporating additional features, and perhaps by taking into account higher-order features. This might help to address the performance gap between our models and human subjects that has been documented in the literature.[3] A more ambitious idea would be to use a statistical model to rank overall parse quality for entire sentences. This would be an improvement over schemes that assume independence between a number of individual scoring functions, such as (Alshawi and Carter, 1994). If such a model were to include only a few general variables to account for such features as lexical association and recency preference for syntactic attachment, it might even be worthwhile to investigate it as an approximation to the human parsing mechanism.

References

Agresti, Alan. 1990. *Categorical Data Analysis.* John Wiley & Sons, New York.

Alshawi, Hiyan and David Carter. 1994. Training and scaling preference functions for disambiguation. *Computational Linguistics*, 20(4):635–648.

Bishop, Y. M., S. E. Fienberg, and P. W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice.* MIT Press, Cambridge, MA.

Charniak, Eugene, Curtis Hendrickson, Neil Jacobson, and Mike Perkowitz. 1993. Equations for part-of-speech tagging. In *AAAI-93*, pages 784–789.

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Crain, Stephen and Mark J. Steedman. 1985. On not being led up the garden path: The use of context by the psychological syntax processor. In David R. Dowty, Lauri Karttunen, and Arnold M. Zwicky, editors, *Natural Language Parsing*, pages 320–358, Cambridge, UK. Cambridge University Press.

de Marcken, Carl G. 1990. Parsing the LOB corpus. In *Proceedings of ACL-90*, pages 243–251.

Deming, W. E. and F. F. Stephan. 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statis*, (11):427–444.

Duda, Richard O. and Peter E. Hart. 1973. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York.

Fienberg, Stephen E. 1980. *The Analysis of Cross-Classified Categorical Data.* The MIT Press, Cambridge, MA, second edition edition.

Franz, Alexander. 1996. *Automatic Ambiguity Resolution in Natural Language Processing.* volume 1171 of *Lecture Notes in Artificial Intelligence.* Springer Verlag, Berlin.

Gibson, Ted and Neal Pearlmutter. 1994. A corpus-based analysis of psycholinguistic constraints on PP attachment. In Charles Clifton Jr., Lyn Frazier, and Keith Rayner, editors, *Perspectives on Sentence Processing.* Lawrence Erlbaum Associates.

Hindle, Donald and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Hirst, Graeme. 1986. *Semantic Interpretation and the Resolution of Ambiguity.* Cambridge University Press, Cambridge.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Ratnaparkhi, Adwait, Jeff Rynar, and Salim Roukos. 1994. A maximum entropy model for Prepositional Phrase attachment. In *ARPA Workshop on Human Language Technology.* Plainsboro, NJ, March 8-11.

Weischedel, Ralph, Marie Meteer, Richard Schwartz, Lance Ramshaw, and Jeff Palmucci. 1993. Coping with ambiguity and unknown words through probabilistic models. *Computational Linguistics*, 19(2):359–382.

[3] For example, If random sentences with "Verb NP PP" cases from the Penn treebank are taken as the gold standard, then (Hindle and Rooth, 1993) and (Ratnaparkhi, Rynar, and Roukos, 1994) report that human experts using only head words obtain 85%-88% accuracy. If the human experts are allowed to consult the whole sentence, their accuracy judged against random Treebank sentences rises to approximately 93%.