

# Actuality and its Correlates

Phillip Potamites

October 31, 2007

## 1 Introduction

This paper surveys statistical distributions of embedded clause entailments. It evaluates the success of algorithms which predict these entailments, and considers aspects of these algorithms which affect performance. It further considers the ability of these algorithms to harvest items which select these entailments, and measures correlations of entailments with other features of embedded clauses.

The data consisted of 500 sentences extracted from *Moby Dick*, and the algorithms were tested on another 50 sentences from the same book. All quotes are from *Moby Dick* unless otherwise noted.

The accuracy of the algorithms, with the best parameters, improved from under 40% to almost 70%, as they trained over these 500 sentences. These scores were well above the base-line, 'best-label' scores of about 35%. Still, the top-score of 69% is hardly competitive by contemporary, computational standards. However, given the minimal assumptions of the bagged observations used for these predictions, and the extremely small size of the available data, this learning progress appears promising. Larger data sets and more structured observations should both be expected to offer enhanced performance. However, this study suggests that learning significant semantic implications on small data sets, with unstructured observations is possible. The limits of these methods also need to be eval-

uated against a yet-unanswered question of human consistency.

Given the small size of the data set, it is also not too surprising that the top-ranked features contained many seemingly irrelevant items. However, of the top 20 mutual information scores for hypotheticality, 50% were *intuitively* related to hypothetical clauses. These significant, preceding features included items like "seem", "tell", "order", "trying", "think", etc.. 40% of the Bayesian scores for actuality were *intuitively* related to actuality; this list included items like "begin" and "knows". Thus, it seems reasonable to hope that larger data sets, with simple interpretative labels, could derive important lexical knowledge about the items that select those labels.

The data also show significant correlations of actuality with finiteness and non-actuality with non-finiteness. Furthermore, among 'that'-clauses, relative clauses are significantly more actual, while complement clauses are significantly more hypothetical. More sophisticated labeling strategies could be directed towards heightening these correlations, perhaps eventually fulfilling the linguist's ideal of absolute determinacy.<sup>1</sup>

---

<sup>1</sup>For instance, it seems likely that the relative clause actuality would depend on the actuality of the NP involved, and we might be able to prove that hypothetical relative clauses always modify imaginary NPs. If correct, we could then argue a perfect correlation, though we'd still be left to determine which NPs are presupposed and which are imagined.

## 2 Annotations

A predicate like ‘dare’ shows how embedded clauses can vary in implication of actuality, as well as subject manifestation.

(1) Implicatives: (see Karttunen (1970), Karttunen (1971))

- a. He dared to speak to the king.  
→ He spoke to the king.
- b. He didn’t dare speak to the king.  
→ He didn’t speak to the king.

(2) a. He dared her to speak to the king.  
→ She spoke to the king.

- b. He didn’t dare her to speak to the king.  
→ She didn’t speak to the king.

Karttunen also points out how some predicates reverse the entailments.

- (3) a. He failed to leave.  
→ He didn’t leave.
- b. He didn’t fail to leave  
→ He left.

Other predicates sustain the entailment under negation.

(4) Factives: (see Kiparsky and Kiparsky (1971))

- a. She was happy that he listened.  
→ He listened.
- b. She was not happy that he left.  
→ He left.

This study collected sentences containing ‘that’, ‘for’, or ‘to’. They first had

to be classified as demonstratives, prepositions, or clauses. ‘That’-clauses were additionally distinguished as relative or complement clauses. ‘For’-clauses were instead categorized as finite (a bit like ‘because’), or overt-subject ‘to’ clauses (the usual form discussed in the linguistics literature), or *other* non-finite forms (frequently ‘-ing’). ‘To’-clauses were distinguished by their subject manifestation (whether immediate, elsewhere in the sentence, or entirely absent from the sentence). The full sets of possible labels, for each kind of cue, are listed below, and then discussed further.

- ‘that’

- [thing=], [CompleteCl:], [RelCl:]
- [actual:], [iffy:], [denied:], []

- ‘for’

- [prep/part:], [for+fin:], [for+subj+to:], [for-fin:]
- [actual:], [iffy:], [denied:], []

- ‘to’

- [prep/part:], [<-above=subj], [=subj], [subj??], [subj->], [to+ing:]
- [actual:], [iffy:], [denied:], []

- ★ Pre-determined dependencies:

- [prep/part:] → []
- [thing=] → []

As noted above, relative modification and finiteness both showed significant correlations with actuality. In contrast, subject

manifestation was not found to have any significant correlation with actuality.

These labels have been chosen with the primary intention of being intuitively obvious to untrained informants. In the future, it would be useful to expand this study by evaluating consistency among numerous informants, so theoretical jargon has been avoided.

Thus, rather than using a label like 'D', as in 'demonstrative', this study uses the label [thing=]. As shown by the 'pre-determined dependencies', this label is necessarily followed by a null label [], and is exempted from questions about its actuality. For instance,

- (5) He's a raal<sup>2</sup> oil-butt , [thing=] [] that fellow ! (-quoting *Cooper's Pilot*)

However, the prepositional category is annotated as [prep/part:], so as to avoid the complex question of distinguishing particles from prepositions. They are equally exempted from actuality questions. For instance, it is unclear if the 'to' in 'belongest to' should be analyzed identically to the 'to' in 'send to', but, for the purposes of this study, they identically select noun phrases, and are therefore categorized as [prep/part:]:

- (6) a. Thou belongest [prep/part:] [] to [thing=] [] that hopeless , sallow tribe which no wine of this world will ever warm ;

<sup>2</sup>I suspect this word to be an old, perhaps pre-standardized, spelling of 'real', or perhaps intended to imitate some vernacular. Other versions of *Moby Dick* maintain the spelling, so it is apparently *not* just a typo of the Gutenberg edition.

- b. " If you make the least damn bit of noise , " replied Samuel , " I will send you [prep/part:] [] to hell . " (-quoting *Life of Samuel Comstock (The Mutineer)*)

This ethic of overt discernibility also blurs a number of distinctions which trained linguists have found significant. The subject categories only hint in what direction to look for the subject, if present at all (while using [subj???], if the subject is completely absent from the sentence). The [<-above=subj] label makes no effort to distinguish raising complements from control complements, nor from purpose clauses:

- (7) They seemed [<-above=subj] [iffy:] to endeavor [<-above=subj] [iffy:] to conceal themselves behind the whale , in order [<-above=subj] [iffy:] to avoid being seen by us . (-quoting *Cook's Voyages*)

Control and raising have been distinguished, by linguists, because a controller (like 'endeavor') assigns a thematic role, whereas no thematic role is assigned by a raiser (like 'seem'). Purpose clauses (like 'in order to', or 'to' used equivalently) have also been distinguished because the control is not obligatory and their subject may just as well be characterized by [subj???] , [=subj], or [subj->].<sup>3</sup>

Identifying these additional categories, or ones like them, may prove useful in subsequently predicting important labels.<sup>4</sup> While purpose interpretations may

<sup>3</sup>Though in what contexts remains a matter of debate.

<sup>4</sup>Jerry Hobbs (p.c.) suggests that identifying silent subjects in adjuncts with the main clause subjects

be straight-forwardly *intuitive*, the control-raising distinction can be controversial, and difficult for untrained informants, or even linguists (see Perlmutter (1970) on the double nature of ‘begin’). This preliminary study, however, only deployed the immediately relevant interpretations and overtly observable properties.

Thus the properties of the overt string are favored over any complications arising from structural assumptions. For instance, non-c-commanding binding has required special theoretical mechanisms, but this study also blends such items into controlled labels.

- (8) a. The voyages of the Dutch and English [prep/part:] [] to the Northern Ocean , in order , if possible , [<-above=subj] [iffy:] to discover a passage through it [prep/part:] [] to India , though they failed of their main object , laid-open the haunts of the whale. (-quoting *McCulloch’s Commercial Dictionary*)

Notice how the above purpose clause implies control from the prepositional phrase of the subject, which structurally could be presumed to be embedded at an unavailable level.<sup>5</sup> While theorists may suggest that this control relation is distinct from that

would probably give better than 80% accuracy. A selection of 50 sentences from *Moby Dick* showed that 86% of the adjunct subjects were identified with the main clause subject. On the other hand, only about 69% of complement subjects were identified with main subjects; As Jerry notes, lexical information can be used to achieve better accuracy for the complements.

<sup>5</sup>Binding theory required the extensions of Higginbotham (1983) or Safir (1984) to cope with a clas-

in a complement relation, it is intuitively clear that it was the Dutch and the English who wished to discover a passage to India. Thus, this current study only queries the intuitive knowledge that the understood subject of the infinitive has been previously mentioned, or not, or will be, etc.

This extreme attempt to refrain from disputable structural decisions has still run into some difficult questions. [=subj] is intended to indicate overt, adjacent subjects, but interveners like negation even appear in main clauses. It seemed unreasonable to suggest that the subject was in some upper, preceding position for every intervener, even if the intervener provided no opportunity of confusing the subject reference. Thus, [=subj] indicates a manifest, adjacent subject in (9a), even though ‘not’ is actually intervening, but the parentheticals in (9b,c), which include NP arguments, are lumped with controlled structures.

- (9) a. So he makes the best of it ; and when the sailors find him not [=subj] [denied:] to be the man [RelCl:] [actual:] that is advertised , they let him pass , and he descends into the cabin .  
 b. I told him [prep/part:] [] for heaven’s sake [<-above=subj] [iffy:] to quit.  
 c. Stammering out something , I knew not what , I rolled away from him against the wall , and then conjured him , whoever

sic sentence like “Some man from every city likes it”. There is a bound reading available for the pronoun, ‘it’, even though the binder, ‘every’, is embedded within the NP.

or whatever he might be , [<-above=subj] [iffy:] to keep quiet , and let me get up and light the lamp again .

As shown and partially discussed, both the complete subject resolution and the proper scope of the various labels are left to future research and/or independent modules.

### 3 Algorithm Comparison

Determining scope requires trustworthy parsing which requires theoretical certainty. Instead word bags divided into preceding and following have been used. This method avoids theoretical assumptions, disputable corpora, and the scarcity of the data. More research is needed to compare the benefits of N-grams, word bags, distance weighting, similarity extensions, and tree-geometric features.

This current study asks how well significant linguistic knowledge can be culled from small sets of unstructured sequential strings, with rudimentary interpretative labels. Parsing has been refrained from because of its dubious theoretical status, and the questionable state of current systems' accuracy, particularly in regard to prepositional and other adjunct attachments. Even N-grams have been dispensed with, due to the problem of 'the sparseness of the data' (also known as 'the poverty of stimulus'). N-grams fix word-features into specific positions. Bags allow a word to be identified with a single feature regardless of its exact sequential positioning. This study attempts to provide evidence that syntactically and semantically significant informa-

tion can arise out of simple statistical precedence and succession.

Existing parsers may show much higher accuracy distinguishing Ps, Ds, and C/Ts. If so, relying on judgments from those systems should remove a large number of errors in assigning the subsequent actuality labels. Future improvements may well make use of more structural information. This study only investigated eliciting label indicators from a small set of unstructured strings. The general correlations described in section 4 rely on these unstructured observations, so it is important to recognize their potentially marginal predictive value.

Labels were predicted, and features ranked, by a variety of evaluation metrics. 3 of these methods are compared below. Features were word occurrences (preceding or following) that were harvested from the training data and counted, where they appeared within a sentence including, or within a given range of, the cues 'that', 'for', or 'to'.

#### 3.1 Bayes

A standard method of predicting labels derives probabilities for any label, given a set of features ( $p(\ell|f_1, \dots, f_n)$ ), from a *naive* application of Bayes rule.

$$p(\ell|f_1, \dots, f_n) = \frac{p(\ell) \prod_{i=0}^n p(f_i|\ell)}{\prod_{i=0}^n p(f_i)} \quad (1)$$

The *naive* assumption lies in both of the substitutions implicit in the above formula.

$$p(f_1, \dots, f_n|\ell) = \prod_{i=0}^n p(f_i|\ell) \quad (2)$$

$$p(f_1, \dots, f_n) = \prod_{i=0}^n p(f_i) \quad (3)$$

Taking the feature product alleviates the data scarcity problem, by providing a way of distinguishing probabilities of unseen joint events.

To avoid zeroing the above equation, and accept the possibility of hitherto unencountered events, “smoothing” redistributes some probability from previously encountered types. As implemented here, we basically pretend that we have seen every possible joint event at least once. The relevant equations, in terms of counts, are given below:<sup>6</sup>

$$p(\ell) = \frac{c(\ell) + |F|}{N + |FL|} \quad (5)$$

$$p(f_i|\ell) = \frac{c(f_i, \ell) + 1}{c(\ell) + |F|} \quad (6)$$

$$p(f_i) = \frac{c(f_i) + |L|}{N + |FL|} \quad (7)$$

$|F|$  is the number of feature *types*,  $|L|$  is the number of label *types*, and  $|FL|$  is their product.

<sup>6</sup>These equations insure the following equalities which are crucial to Bayes’ Theorem:

$$\frac{p(\ell)p(f|\ell)}{p(f)} = \frac{p(f, \ell)}{p(f)} = p(\ell|f) \quad (4)$$

It is also important that  $N=c(L)=c(F)$ , where  $c(L)$  and  $c(F)$  are the counts of all label and feature events respectively. The naive/independence assumption considers every feature-label co-occurrence a single event, so labels are counted for exactly the number of features that they possess. An analysis that only counted labels for their proper occurrence would seem to require a theory of Bayesian Expectation.

### 3.2 Mutual Information

Mutual information is another way of evaluating probability relationships.<sup>7</sup> It is normally given by the following formula.

$$mi_{\ell f} = p(\ell, f) \log \frac{p(\ell, f)}{p(\ell)p(f)} \quad (8)$$

Applying this metric to predict labels, it seemed reasonable to sum the contributions of features present in a given observation which we wish to label. The logarithmic scale however makes some contributions less than 0, where 0 seemed a more preferable baseline for no known evidence. So it was discarded. Maintaining the joint probability coefficient, with the result of squaring the nominator, actually hurt performance slightly. On the other hand, giving some weight to the number of features active in a given observation, for a given label, made for significant improvement.

Because mutual information tends to overvalue the importance of rare events, performance is improved by discounting by a ratio of either the label or feature count, whichever is smaller. The final equations, in terms of counts, are shown below.

$$\ell = \arg \max_{\ell \in \mathcal{L}} n \sum_{i=1}^n dN \frac{c(\ell, f_i)}{c(\ell)c(f_i)} \quad (9)$$

$$d = \frac{\min(c(\ell), c(f_i))}{\min(c(\ell), c(f_i)) + 1} \quad (10)$$

$N$  is the count of all labeled events, which is the remaining normalizer from the probability equation above (being the same for all

<sup>7</sup>Thanks to Patrick Pantel for his classroom material presenting mutual information.

observations, it can be ignored).  $d$  is the discounting factor.  $n$  is just the number of features active for that potential labeling event. Basically, it makes 2 mediocre features better than 1 good one.

### 3.3 Best Label

The above 2 evaluation metrics can be compared to a base-line method of prediction. The ‘best label’ method just calculates the label that is most likely to appear in any position for any cue, and always assigns that label. In this study, that meant always assigning [prep/part:] for any appearance of ‘for’ or ‘to’, and alternating between [thing=] and [CompleteCl:] for ‘that’. Because the test set had a lower proportion of nominal ‘that’s, the [thing=] method only made 33%, whereas the [CompleteCl:] strategy scored 37%.<sup>8</sup>

### 3.4 Results

Table 1 shows the performance of each method, in predicting ‘[prep/part:]’, using bags of 3 preceding and 3 following words. The mi method is far more conservative in deploying the most likely [prep/part:] label, but its overall f-measure still turns out exactly the same. The bayes method over-deploys this label, but at least the recall benefits from this optimism. In fact, the prepositional probability was much higher in the training set than in the test set. For instance,

<sup>8</sup>Here are the relevant counts with percentages:

| ‘that’        | train    | test    |
|---------------|----------|---------|
| [thing=]      | 114(43%) | 7(26%)  |
| [CompleteCl:] | 110(41%) | 15(56%) |
| [RelCl:]      | 44(16%)  | 5(18%)  |

whereas 82% of the ‘for’-occurrences were prepositional in the training set, only 56% were prepositional in the test set. These proportions are also listed in tables 22 and 23.

Table 1: [prep/part:]

| 26 possible answers | bayes | mi  | bl  |
|---------------------|-------|-----|-----|
| proposed answers    | 44    | 24  | 68  |
| correct answers     | 25    | 18  | 26  |
| recall              | .96   | .69 | 1.0 |
| precision           | .57   | .75 | .38 |
| f-measure           | .72   | .72 | .55 |

Table 2: [iffy:]

| 42 possible answers | bayes | mi  | bl  |
|---------------------|-------|-----|-----|
| proposed answers    | 25    | 32  | 0   |
| correct answers     | 22    | 24  | 0   |
| recall              | .52   | .57 | 0   |
| precision           | .88   | .75 | n/a |
| f-measure           | .65   | .65 | n/a |

For ‘[iffy:]’ (table 2), the mi method is more optimistic. Notice that one of the primary differences between the methods lies in multiplying or dividing by the label count/probability. Thus the mi method is more generous with the label that is less likely from a non-conditioned perspective. That optimism did not, however, bring much recall, and the overall f-measure is again the same. As above, significant differences in the two sample distributions also appear influential. [iffy:] was much more likely in the test set than in the training: 22% of ‘that’s were [iffy:] in the training set, and 7% of ‘for’s, compared to 40% of ‘that’s and 22% of ‘for’s, in the test set. See tables 22, 23. These differences partially account for the significant underestimation of the [iffy:] label, by these algorithms.

Tables 3-6 show the overall performance of the different methods as they learn from

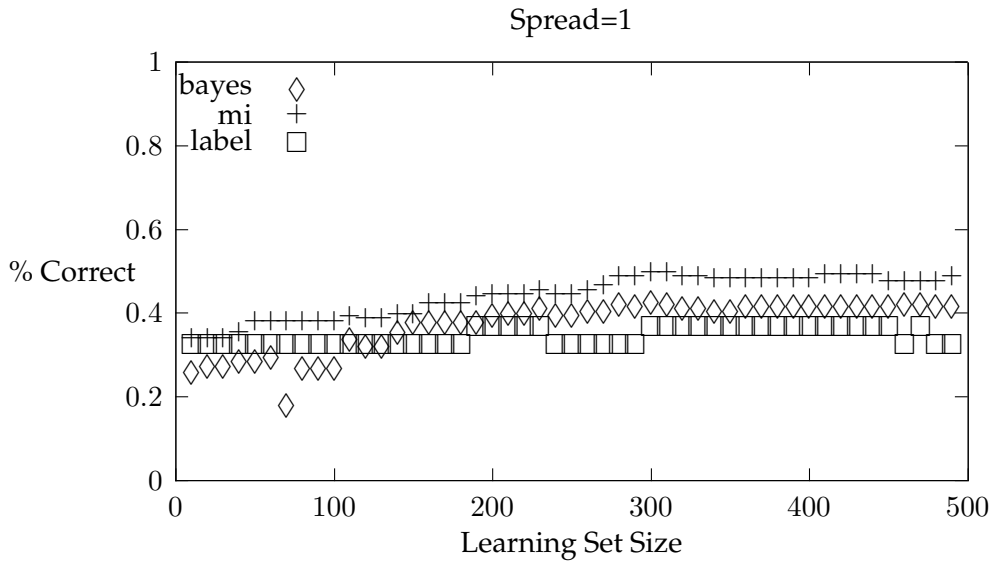


Table 3: Grade by Training

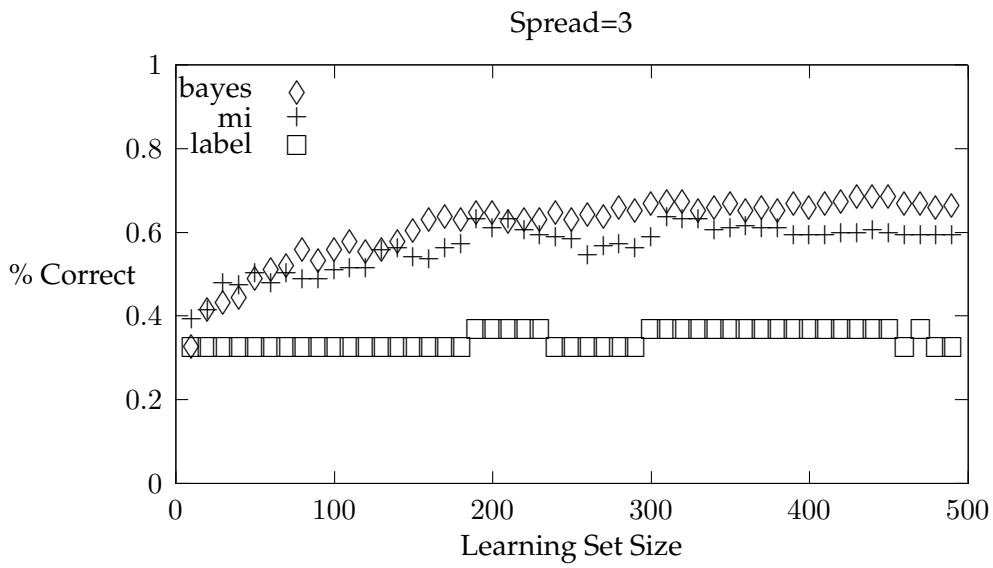


Table 4: Grade by Training

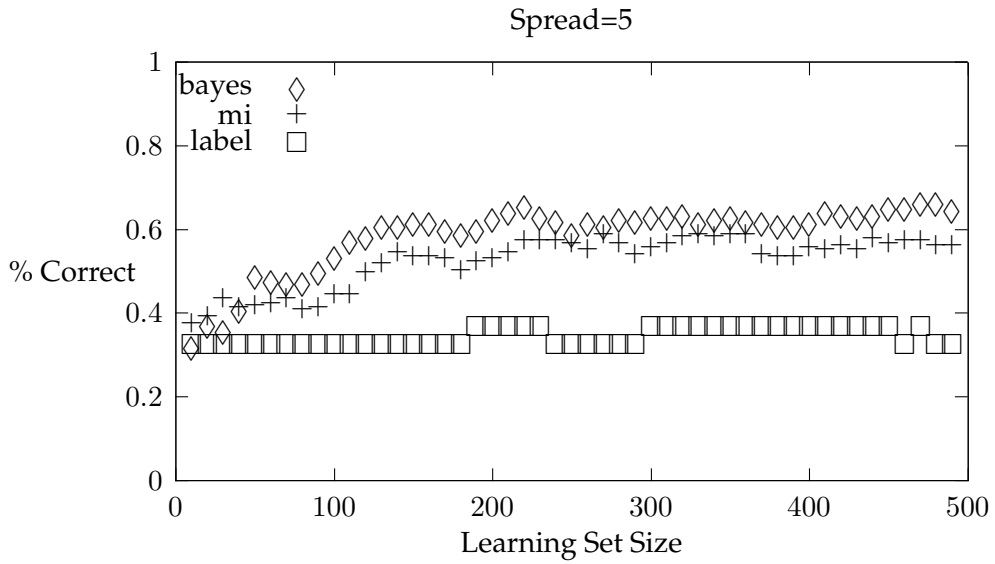


Table 5: Grade by Training

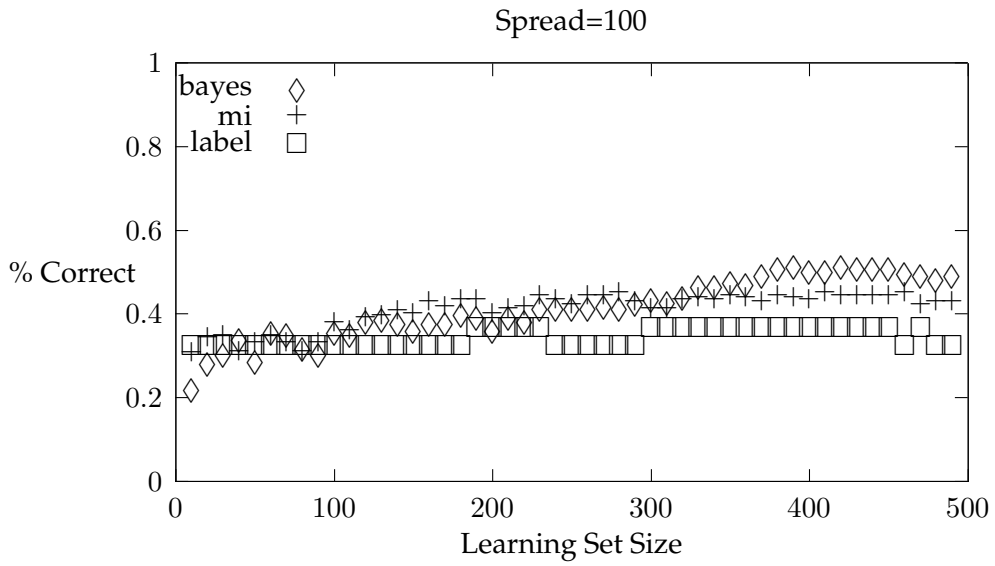


Table 6: Grade by Training

10 to 490 labeled sentences. The different tables correspond to different sized bags, with spread referring to the distance, in words, considered on each side of a cue.

They make a simple point, worth demonstrating: *more is not always better*. In so far as eliciting information from bags is concerned, considering ever more distant features only increases the amount of noise in the system. Future implementations may find suitable ways to weight feature distance. Or structural information, appropriately handled, may begin to determine what kind of distances should be traversed.

Hopefully, larger training sets would continue the improvement, but consistency may be limited further by both observational limitations and subjective fluctuations of human judgments. These are 3 directions for future research.

## 4 Ranks and Correlations

For a single feature, and a given label, the main difference in algorithms lies in the difference between smoothing and discounting. As shown by tables 7 and 8, their top-ranked features, for the most part, overlap. Both methods learn the pervasive importance of “seemed”, “seem”, and “order” (basically from the purpose clauses) for predicting non-actuality, as shown in table 7. They also learn the importance of “began”, “found”, and “maketh”<sup>9</sup> for predicting actuality, as shown in table 8.

However, they also attribute significance

<sup>9</sup>The great Leviathan [RelCl:] [actual:] that maketh the seas [=subj] [actual:] to seethe like a boiling pan . (-quoting Lord Bacon’s Version of the Psalms)

to useless, over-fit, coincidences, like ‘He’, which we intuitively know to be irrelevant. Such distractor features are likely to decrease performance.

Brent (1991) suggests a binomial test, in accepting success for some cued evidence as real evidence of that subcategorization frame. Here, there is no assumption that words are positively or negatively subcategorized. They are just more or less associated.

This study examines whether the classic  $\chi^2$ -test can help distinguish the ‘over-fit’ associations, presumably arisen from our sparse data, and noisy, incomplete representations, from those features which we intuitively know to imply intentional worlds. The goal is for trainable systems to be able to automatically gather this kind of knowledge from strings with elementary interpretative labels.

As shown by tables 9, 10, “seemed” easily passes the  $\chi^2$ -test for statistical significance, but “seem” just barely misses the commonly accepted p-value of  $<.05$ . The failure signifies little more than the value of morphological parsing. With enough data, such knowledge might also arise from statistical combinatorics. There may also be questions about the validity of  $\chi^2$ -tests for such small cell values.

Table 9:  $p(\chi^2 > 7.63, f^\circ = 1) = .0057$

| pre:      | seemed | *   |
|-----------|--------|-----|
| [actual:] | 0      | 171 |
| [iffy:]   | 11     | 242 |

Fortunately, the  $\chi^2$ -test successfully rejects ‘He’, as shown by table 11.

But it also rejects items which were intu-

Table 7: [iffy:] :pre:

|     |        | mi   |           | bayes |
|-----|--------|------|-----------|-------|
| 1.  | seemed | 2.90 | seemed    | 0.8   |
| 2.  | seem   | 2.40 | seem      | 0.67  |
| 3.  | order  | 2.21 | order     | 0.63  |
| 4.  | He     | 2.05 | He        | 0.6   |
| 5.  | would  | 1.94 | would     | 0.57  |
| 6.  | almost | 1.94 | who       | 0.57  |
| 7.  | into   | 1.94 | almost    | 0.57  |
| 8.  | trying | 1.94 | into      | 0.57  |
| 9.  | made   | 1.94 | trying    | 0.57  |
| 10. | who    | 1.92 | made      | 0.57  |
| 11. | do     | 1.84 | do        | 0.56  |
| 12. | than   | 1.84 | than      | 0.56  |
| 13. | are    | 1.84 | are       | 0.56  |
| 14. | man    | 1.80 | man       | 0.55  |
| 15. | take   | 1.80 | take      | 0.55  |
| 16. | what   | 1.76 | what      | 0.53  |
| 17. | have   | 1.60 | concluded | 0.5   |
| 18. | told   | 1.58 | nigh      | 0.5   |
| 19. | tell   | 1.55 | seeks     | 0.5   |
| 20. | think  | 1.55 | never     | 0.5   |

Table 8: [actual:] ;pre:

|     |              | mi   |              | bayes |
|-----|--------------|------|--------------|-------|
| 1.  | began        | 3.25 | began        | 0.63  |
| 2.  | found        | 2.29 | maketh       | 0.5   |
| 3.  | maketh       | 2.26 | sounds       | 0.5   |
| 4.  | sounds       | 2.26 | circumstance | 0.5   |
| 5.  | circumstance | 2.26 | moment       | 0.5   |
| 6.  | moment       | 2.26 | window       | 0.5   |
| 7.  | window       | 2.26 | knows        | 0.5   |
| 8.  | knows        | 2.26 | sea          | 0.5   |
| 9.  | sea          | 2.26 | begin        | 0.5   |
| 10. | begin        | 2.26 | once         | 0.5   |
| 11. | once         | 2.26 | found        | 0.5   |
| 12. | was          | 2.05 | he's         | 0.43  |
| 13. | now          | 1.81 | whales       | 0.43  |
| 14. | the          | 1.80 | was          | 0.43  |
| 15. | he's         | 1.70 | Leviathan    | 0.43  |
| 16. | whales       | 1.70 | well         | 0.43  |
| 17. | Leviathan    | 1.70 | same         | 0.43  |
| 18. | well         | 1.70 | sure         | 0.43  |
| 19. | same         | 1.70 | nothing      | 0.43  |
| 20. | sure         | 1.70 | The          | 0.43  |

Table 10:  $p(\chi^2 > 3.42, f^\circ = 1) = .064$

| pre:      | seem | *   |
|-----------|------|-----|
| [actual:] | 0    | 171 |
| [iffy:]   | 5    | 248 |

Table 11:  $p(\chi^2 > 1.42, f^\circ = 1) = .23$

| pre:      | He | *   |
|-----------|----|-----|
| [actual:] | 1  | 170 |
| [iffy:]   | 5  | 248 |

itively appealing, like “trying” and “told”, as shown by tables 12 and 13, respectively.

Table 12:  $p(\chi^2 > 2.04, f^\circ = 1) = .15$

| pre:      | trying | *   |
|-----------|--------|-----|
| [actual:] | 0      | 171 |
| [iffy:]   | 3      | 250 |

Table 13:  $p(\chi^2 > 2.73, f^\circ = 1) = .099$

| pre:      | told | *   |
|-----------|------|-----|
| [actual:] | 0    | 171 |
| [iffy:]   | 4    | 249 |

It is possible to attribute this failure to the lack of data. Given the way that the  $\chi^2$ -test works, if we doubled our observations, as supposed in table 14, even ‘trying’ would pass. On the other hand, given its outside co-occurrence with [actual:], such supposition, as imagined in table 15, would not help ‘He’ beat the arbitrary standard of 5%, though it starts coming close.

Table 14:  $p(\chi^2 > 4.08, f^\circ = 1) = .043$

|   |     |
|---|-----|
| 0 | 342 |
| 6 | 500 |

Table 15:  $p(\chi^2 > 2.83, f^\circ = 1) = .092$

|    |     |
|----|-----|
| 2  | 340 |
| 10 | 496 |

Surprisingly, ‘who’ really does seem, in this data, to correlate with non-actuality, as presented in table 16.<sup>10</sup> The association seems particularly bizarre in that ‘who’ is often associated with relative clauses, which will here be shown, in ‘that’-clauses, to positively correlate with actuality.

Table 16:  $p(\chi^2 > 4.78, f^\circ = 1) = .029$

| pre:      | who | *   |
|-----------|-----|-----|
| [actual:] | 0   | 170 |
| [iffy:]   | 7   | 246 |

The correlation arises from phrases like the following:

- (10) a. This savage was the only person present who seemed [ $\leftarrow$ -above=subj] [iffy:] to notice my entrance .
- b. Woe [prep/part:] [] to him who seeks [ $\leftarrow$ -above=subj] [iffy:] to pour oil upon the waters when God has brewed them into a gale !
- (11) a. a deep stupor steals over him , as over the man who bleeds [ $\leftarrow$ -above=subj] [iffy:] to death
- b. And eternal delight and deliciousness will be his , who coming [ $\leftarrow$ -above=subj] [iffy:] to lay him

<sup>10</sup>Note that this issue also arises particularly because features were only collected where they were associated with labels.

down , can say with his final  
breath – O Father !

Whereas ‘who’ in examples (10a,b) might be said to ‘accidentally’ embed an intensional selector, the ‘who’ in examples (11a,b) actually seems to transmit the hypothetical nature of the head noun. This kind of usage suggests a possible tendency to use ‘who’ rather than ‘that’ with more hypothetical entities. Such speculation is another avenue for future research.

Thus, the  $\chi^2$ -test can help exclude some rash conclusions; however, depending on whether we accept ‘who’ as appropriately correlated, some over-fit coincidences may be sustained by this limited data.

When other labels were included in the feature rankings, labels tended to fill in the top ranked spots (by design, labels came in pairs). Table 17 shows the top-ranked preceding features for the [iffy:] label, *with* other labels, and table 18 does the same for [actual:].

Among these labels, a pattern seems to emerge. The labels associated with [iffy:] also tend to be those associated with non-finiteness, while the labels associated with [actual:] are also more associated with finiteness. We can test this correlation with a  $\chi^2$ -test, as well, and as table 19 shows, it handily defies the expectations of the null hypothesis.

[CompleteCl:] also appears under both the [iffy:] and [actual:] rankings (tables 17, 18), while [RelCl:] only appears with [actual:] (table 18). A  $\chi^2$ -test suggests that there is a significant correlation between actuality and relativity (table 20). Increasing the correlation probably would involve more precise labeling of the argu-

Table 19:  $p(\chi^2 > 76.72, f^\circ = 1) = 1.97 \times 10^{-18}$

|   | finite | non-finite |
|---|--------|------------|
| [actual:]                                   | 114    | 60         |
| [iffy:]                                     | 58     | 193        |
| finite: [CompleteCl:], [RelCl:], [for+fin:] |        |            |
| non-finite: [for-fin:], [for+subj+to:]      |        |            |
| [<-above=subj],[=subj],[subj??],[subj->]    |        |            |

ments involved, and a more developed theory of imaginary NPs, and intentional contexts. The rudimentary labeling strategy employed here only begins to point in directions for more precise inquiry.

Table 20:  $p(\chi^2 > 5.83, f^\circ = 1) = .0157$

|           | [CompleteCl:] | [RelCl:] |
|-----------|---------------|----------|
| [actual:] | 65            | 33       |
| [iffy:]   | 48            | 9        |

In contrast, the remarkable behavior displayed by the alternations of ‘dare’, at the beginning of this paper, does not appear to be part of any generalizable correlation (table 21). Capturing this correlation would presumably require more lexically specific information.

Table 21:  $p(\chi^2 > 2.63, f^\circ = 3) = .4514$

|           | [<-] | [=] | [??] | [->] |
|-----------|------|-----|------|------|
| [actual:] | 48   | 7   | 3    | 2    |
| [iffy:]   | 134  | 30  | 19   | 4    |

Linguists have proposed correlations among subject interpretations and tense (Landau (2000)), and among ‘for’, subjects, and irrealis interpretations (Pesetsky and Torrego (2001)-(2006)). This study took a more general view of subject manifestation

Table 17: [iffy:] :pre:

|     |                | mi   |                | bayes |
|-----|----------------|------|----------------|-------|
| 1.  | [subj???       | 4.96 | [<-above=subj] | 0.67  |
| 2.  | [=subj]        | 4.89 | [=subj]        | 0.55  |
| 3.  | [<-above=subj] | 4.88 | [subj???       | 0.5   |
| 4.  | [for+subj+to:] | 4.40 | [CompleteCl:]  | 0.38  |
| 5.  | [to+ing:]      | 4.09 | seemed         | 0.30  |
| 6.  | [for-fin:]     | 3.06 | [to+ing:]      | 0.27  |
| 7.  | seemed         | 2.84 | [for+subj+to:] | 0.25  |
| 8.  | [CompleteCl:]  | 2.84 | seem           | 0.23  |
| 9.  | [subj->]       | 2.75 | who            | 0.23  |
| 10. | seem           | 2.60 | [subj->]       | 0.22  |

Table 18: [actual:] :pre:

|     |               | mi   |                | bayes |
|-----|---------------|------|----------------|-------|
| 1.  | [for+fin:]    | 8.05 | [RelCl:]       | 0.58  |
| 2.  | [RelCl:]      | 7.39 | [for+fin:]     | 0.5   |
| 3.  | [CompleteCl:] | 5.43 | [CompleteCl:]  | 0.49  |
| 4.  | began         | 3.61 | [<-above=subj] | 0.24  |
| 5.  | found         | 2.86 | began          | 0.21  |
| 6.  | circumstance  | 2.71 | found          | 0.17  |
| 7.  | knows         | 2.71 | was            | 0.17  |
| 8.  | sea           | 2.71 | [=subj]        | 0.15  |
| 9.  | begin         | 2.71 | circumstance   | 0.15  |
| 10. | once          | 2.71 | knows          | 0.15  |

and hypotheticality, and did not directly test those claims, so the lack of correlation displayed here does not directly discount the claims made there. Hopefully future studies will adhere directly to their labeling claims, and test the statistical significance of the correlations they posit.

## 5 %

For the sake of general summary, table 22 presents the percentage distribution of the 3 relevant functional items as D/Ps or clause markers associated with different implications, across the sample of 500 sentences studied here.

Table 22: Training Set Dist.

|              | that | for | to  |
|--------------|------|-----|-----|
| [thing=] ,   |      |     |     |
| [prep/part:] | .43  | .82 | .46 |
| [actual:]    | .35  | .11 | .12 |
| [iffy:]      | .216 | .07 | .39 |
| [denied:]    | .004 | 0   | .02 |

The distributions of the test set are also provided for comparison (table 23).

Table 23: Test Set Dist.

|              | that | for | to  |
|--------------|------|-----|-----|
| [thing=] ,   |      |     |     |
| [prep/part:] | .26  | .56 | .39 |
| [actual:]    | .29  | .22 | .11 |
| [iffy:]      | .41  | .22 | .47 |
| [denied:]    | .04  | 0   | .03 |

## 6 Future Research

Numerous developments have suggested themselves. More data is called for. New labeling strategies can be entertained. More elaborate feature observations should be developed. New cue sets should be collected.

This study relies completely on the presence of key function words, but clauses can be introduced by many predicates, often without the overt use of these function words. Thus, future studies will expand the set of cues to collect the selectional distributions of many such predicates. Cues like “who” or “how” would also expand the data relevant to complement/relative clause interpretations.

Possible labeling alternatives have also been suggested along the way. Ad-junct/complement distinctions could correlate well with subject identification, as suggested by Jerry Hobbs (see fn. 4). The proposals of Pesetsky and Torrego (2001) or Landau (2000) also imply other relevant properties of clauses, though assigning them would also require more subtle linguistic judgments.

It also might be appropriate to use graded judgments for these kinds of interpretations. It is almost ironic that this paper makes so much of probabilistic inference, and yet still assigns coarse categories of certainty like [actual:]. This matter of actuality seems highly susceptible to an analysis that attributes degrees of likelihood. How to deploy, interpret, and compare such scalar judgments is also a topic for future research.

This study also made use of the coarsest possible observations: unordered word bags. More sophisticated observations need

to be compared. N-grams have decent track records, but tree-geometric relations seem essential to uncovering real causal relations. Intervention may also prove to be an important feature in predicting entailments.

Human consistency for hypothesized, interpretative labeling strategies also must be compared. Historical texts could also be investigated, with similar quantitative measures, to compare linguistic overlap, and demonstrate dialect differences.

## References

- Bahl, Lalit R., Brown, Peter F., de Souza, Peter V., and Mercer, Robert L. (1986) "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," in *ICASSP*, pp. 231–234.
- Baird, Davis (1983) "The Fisher/Pearson Chi-Squared Controversy: A Turning Point for Inductive Inference," *Br J Philos Sci*, 34(2), 105–118.
- Berger, Adam L., Pietra, Stephen Della, and Pietra, Vincent J. Della (1996) "A Maximum Entropy Approach to Natural Language Processing," *Computational Linguistics*, 22(1), 39–71.
- Brent, Michael R. (1991) "Automatic Acquisition of Subcategorization Frames from Untagged Text," in *Meeting of the Association for Computational Linguistics*, pp. 209–214.
- Brown, Peter F., Cocke, John, Pietra, Stephen Della, Pietra, Vincent J. Della, Jelinek, Frederick, Lafferty, John D., Mercer, Robert L., and Roossin, Paul S. (1990) "A Statistical Approach to Machine Translation," *Computational Linguistics*, 16(2), 79–85.
- Darroch, J.N. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models," *Ann. Math. Statist.*, 43, 1470–1480.
- Higginbotham, James (1983) "Logical Form, Binding, and Nominals," *Linguistic Inquiry*, 14, 395–420.
- Karttunen, Lauri (1970) "On the semantics of complement sentences," in *CLS 6*, pp. 328–339.
- Karttunen, Lauri (1971) "Implicative Verbs," *Language*, 47, 340–358.
- Kiparsky, Paul and Kiparsky, Carol (1971) "Fact," in Steinberg, D. and Jakobovits, L., eds., *Semantics*, pp. 143–173, Cambridge.
- Landau, Idan (2000) *Elements of Control*, Kluwer, Dordrecht.
- Malouf, R. (2002) "A comparison of algorithms for maximum entropy parameter estimation," .
- Melville, Herman (1851) "Moby Dick," <http://www.gutenberg.org/etext/2701>.
- Perlmutter, David M. (1970) "The two verbs begin," in *Readings in transformational grammar*, pp. 107–19, Ginn and Co., Waltham, MA.
- Pesetsky, David and Torrego, Esther (2001) "T-to-C Movement: Causes and Consequences," [Final version in Michael Kenstowicz (ed.), *Ken Hale: A Life on*

*Language*. Cambridge, MA: MIT Press. (2001)].

Pietra, Stephen Della, Pietra, Vincent J. Della, and Lafferty, John D. (1997) "Inducing Features of Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4), 380–393.

Ratnaparkhi, A. (1998) "Maximum Entropy Models for Natural Language Ambiguity Resolution," Ph.D Thesis, University of Pennsylvania.

Safir, Ken (1984) "Multiple Variable Binding," *Linguistic Inquiry*, 15(4), 603–638.

Smith, Noah A. (2004) "Log-Linear Models," Johns Hopkins, [nasmith@cs.jhu.edu](mailto:nasmith@cs.jhu.edu).