Modeling Human Sentence Processing Data with a Statistical Parts-of-Speech Tagger

Jihyun Park
Department of Linguisitcs
The Ohio State University
Columbus, OH, USA
park@ling.ohio-state.edu

Abstract

It has previously been assumed in the psycholinguistic literature that finite-state models of language are crucially limited in their explanatory power by the locality of the probability distribution and the narrow scope of information used by the model. We show that a simple computational model (a bigram part-of-speech tagger based on the design used by Corley and Crocker (2000)) makes correct predictions on processing difficulty observed in a wide range of empirical sentence processing data. We use two modes of evaluation: one that relies on comparison with a control sentence, paralleling practice in human studies; another that measures probability drop in the disambiguating region of the sentence. Both are surprisingly good indicators of the processing difficulty of garden-path sentences. The sentences tested are drawn from published sources and systematically explore five different types of ambiguity: previous studies have been narrower in scope and smaller in scale. We do not deny the limitations of finite-state models, but argue that our results show that their usefulness has been underestimated.

1 Introduction

The main purpose of the current study is to investigate the extent to which a probabilistic part-of-speech (POS) tagger can correctly model human sentence processing data. Syntactically ambiguous sentences have been studied in great depth in psycholinguistics because the pattern of ambiguity resolution provides a window onto the human sentence processing mechanism (HSPM). *Prima facie* it seems unlikely that such a tagger will be adequate, because almost all previous researchers have assumed, following standard linguistic theory, that a formally adequate account of recursive syntactic structure is an essential component of any model of the behaviour. In this study, we tested a bigram POS tagger on different types of structural ambiguities and (as a sanity check) to the well-known asymmetry of subject and object relative clause processing.

Theoretically, the garden-path effect is defined as processing difficulty caused by reanalysis. Empirically, it is attested as comparatively slower reading time or longer eye fixation at a disambiguating region in an ambiguous sentence compared to its control sentences (Frazier and Rayner, 1982; Trueswell, 1996). That is, the garden-path effect detected in many human studies, in fact, is measured through a "comparative" method.

This characteristic of the sentence processing research design is reconstructed in the current study using a probabilistic POS tagging system. Under the assumption that larger probability decrease indicates slower reading time, the test results suggest that the probabilistic POS tagging system can predict reading time penalties at the disambiguating region of garden-path sentences compared to that of non-garden-path sentences (i.e. control sentences).

2 Experiments

A Hidden Markov Model POS tagger based on bigrams was used. We made our own implementation to be sure of getting as close as possible to the design of Corley and Crocker (2000). Given a word string, $w_0, w_1, \cdots, w_n$, the tagger calculates the probability of every possible tag path,

Proceedings of the COLING/ACL 2006 Student Research Workshop, pages 25–30,
Sydney, July 2006.

$t_0, \cdots, t_n$. Under the Markov assumption, the joint probability of the given word sequence and each possible POS sequence can be approximated as a product of conditional probability and transition probability as shown in (1).

(1) $P(w_0, w_1, \cdots, w_n, t_0, t_1, \cdots, t_n)$

$\approx \Pi_{i=1}^{n} P(w_i|t_i) \cdot P(t_i|t_{i-1})$, where $n \geq 1$.

Using the Viterbi algorithm (Viterbi, 1967), the tagger finds the most likely POS sequence for a given word string as shown in (2).

(2) $\arg\max P(t_0, t_1, \cdots, t_n|w_0, w_1, \cdots, w_n, \mu)$.

This is known technology, see Manning and Schütze (1999), but the particular use we make of it is unusual. The tagger takes a word string as an input, outputs the most likely POS sequence and the final probability. Additionally, it presents accumulated probability at each word break and probability re-ranking, if any. Probability re-ranking occurs when a previously less preferred POS sequence is more favored later. Note that the running probability at the beginning of a sentence will be 1, and will keep decreasing at each word break since it is a product of conditional probabilities.

We tested the predictability of the model on empirical reading data with the probability decrease and the presence or absence of probability re-ranking. Probability re-ranking occurs when a less preferred POS sequence is selected later over a temporarily favored sequence. Adopting the standard experimental design used in human sentence processing studies, where word-by-word reading time or eye-fixation time is compared between an experimental sentence and its control sentence, this study compares probability at each word break between a pair of sentences. Comparatively faster drop of probability is expected to be a good indicator of comparative processing difficulty. Probability re-ranking, which is a simplified model of the reanalysis process assumed in many human studies, is also tested as another indicator of garden-path effect. Probability re-ranking will occur when an initially dispreferred POS subsequence becomes the preferred candidate later in the parse, because it fits in better with later words.

The model parameters, $P(w_i|t_i)$ and $P(t_i|t_{i-1})$, are estimated from a small section (970,995 tokens,47,831 distinct words) of the British National Corpus (BNC), which is a 100 million-word collection of British English, both written and spoken, developed by Oxford University Press (Burnard, 1995). The BNC was chosen for training the model because it is a POS-annotated corpus, which allows supervised training. In the implementation we use log probabilities to avoid underflow, and we report log probabilities in the sequel.

2.1 Hypotheses

If the HSPM is affected by frequency information, we can assume that it will be easier to process events with higher frequency or probability compared to those with lower frequency or probability. Under this general assumption, the overall difficulty of a sentence is expected to be measured or predicted by the mean size of probability decrease. That is, probability will drop faster in garden-path sentences than in control sentences (e.g. unambiguous sentences or ambiguous but non-garden-path sentences).

More importantly, the probability decrease pattern at disambiguating regions will predict the trends in the reading time data. All other things being equal, we might expect a reading time penalty for a garden-path region when the size of the probability decrease at the disambiguating region of a garden-path sentence will be greater than that of control sentences. This is a simple and intuitive assumption that can be easily tested. We could have formed the sum over all possible POS sequences in association with the word strings, but for the present study we simply used the Viterbi path: justifying this because this is the best single-path approximation to the joint probability.

Lastly, re-ranking of POS sequences is expected to predict reanalysis of lexical categories. This is because re-ranking in the tagger is parallel to reanalysis in human subjects, which is known to be cognitively costly.

2.2 Materials

In this study, five different types of ambiguity were tested including Lexical Category ambiguity, Reduced-relative ambiguity (RR ambiguity), Preposition-phrase attachment ambiguity (PP ambiguity), Direct-object/Sentential-complement ambiguity (DO/SC ambiguity), and Clausal Boundary ambiguity. The following are example sentences for each ambiguity type, shown with the ambiguous region italicized and the dis-

ambiguating region bolded. All of the example sentences are garden-path sentneces.

(3) Lexical Category ambiguity
The foreman knows that the warehouse *prices* **the** beer very modestly.

(4) RR ambiguity
The horse *raced* past the barn **fell**.

(5) PP ambiguity
Katie laid the dress *on the floor* **onto** the bed.

(6) DO/SC ambiguity
He forgot Pam **needed** a ride with him.

(7) Clausal Boundary ambiguity
Though George kept on reading *the story* really **bothered** him.

The test materials are constructed such that a garden-path sentence and its control sentence share exactly the same word sequence except for the disambiguating word so that extraneous variables such as word frequency effect can be controlled. We inherit this careful design.

In this study, a total of 76 sentences were tested: 10 for lexical category ambiguity, 12 for RR ambiguity, 20 for PP attachment ambiguity, 16 for DO/SC ambiguity, and 18 for clausal boundary ambiguity. This set of materials is, to our knowledge, the most comprehensive yet subjected to this type of study. The sentences are directly adopted from various psycholinguistic studies (Frazier, 1978; Trueswell, 1996; Ferreira and Henderson, 1986).

As a baseline test case of the tagger, the well-established asymmetry between subject- and object-relative clauses was tested as shown in (8).

(8) a. The editor who kicked the writer fired the entire staff. (Subject-relative)
b. The editor who the writer kicked fired the entire staff. (Object-relative)

The reading time advantage of subject-relative clauses over object-relative clauses is robust in English (Traxler et al., 2002) as well as other languages (Mak et al., 2002; Homes et al., 1981). For this test, materials from Traxler et al. (2002) (96 sentences) are used.

3 Results

3.1 The Probability Decrease per Word

Unambiguous sentences are usually longer than garden-path sentences. To compare sentences of different lengths, the joint probability of the whole sentence and tags was divided by the number of words in the sentence. The result showed that the average probability decrease was greater in garden-path sentences compared to their unambiguous control sentences. This indicates that garden-path sentences are more difficult than unambiguous sentences, which is consistent with empirical findings.

Probability decreased faster in object-relative sentences than in subject relatives as predicted. In the psycholinguistics literature, the comparative difficulty of object-relative clauses has been explained in terms of verbal working memory (King and Just, 1991), distance between the gap and the filler (Bever and McElree, 1988), or perspective shifting (MacWhinney, 1982). However, the test results in this study provide a simpler account for the effect. That is, the comparative difficulty of an object-relative clause might be attributed to its less frequent POS sequence. This account is particularly convincing since each pair of sentences in the experiment share the exactly same set of words except their order.

3.2 Probability Decrease at the Disambiguating Region

A total of 30 pairs of a garden-path sentence and its ambiguous, non-garden-path control were tested for a comparison of the probability decrease at the disambiguating region. In 80% of the cases, the probability drops more sharply in garden-path sentences than in control sentences at the critical word. The test results are presented in (9) with the number of test sets for each ambiguous type and the number of cases where the model correctly predicted reading-time penalty of garden-path sentences.

(9) Ambiguity Type (Correct Predictions/Test Sets)
a. Lexical Category Ambiguity (4/4)
b. PP Attachment Ambiguity (10/10)
c. RR Ambiguity (3/4)
d. DO/SC Ambiguity (4/6)
e. Clausal Boundary Ambiguity (3/6)

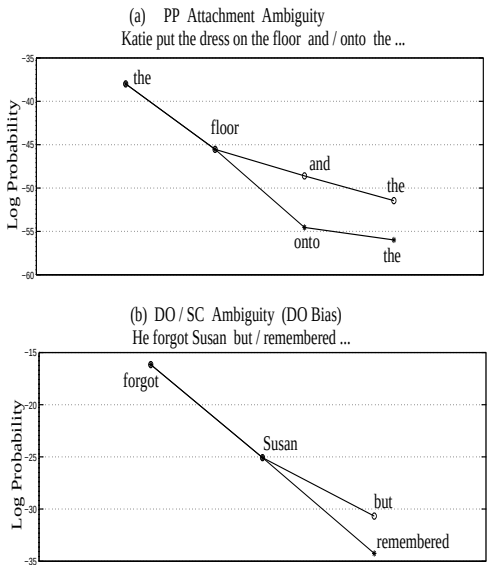

Figure 1: Probability Transition (Garden-Path vs. Non Garden-Path)

(a) $-\circ-$: Non-Garden-Path (Adjunct PP), $-*-$: Garden-Path (Complement PP)
(b) $-\circ-$: Non-Garden-Path (DO-Biased, DO-Resolved), $-*-$: Garden-Path (DO-Biased, SC-Resolved)

The two graphs in Figure 1 illustrate the comparison of probability decrease between a pair of sentence. The *y*-axis of both graphs in Figure 1 is log probability. The first graph compares the probability drop for PP ambiguity (*Katie put the dress on the floor and/onto the bed....*) The empirical result for this type of ambiguity shows that reading time penalty is observed when the second PP, *onto the bed*, is introduced, and there is no such effect for the other sentence. Indeed, the sharper probability drop indicates that the additional PP is less likely, which makes a prediction of a comparative processing difficulty. The second graph exhibits the probability comparison for the DO/SC ambiguity. The verb *forget* is a DO-biased verb and thus processing difficulty is observed when it has a sentential complement. Again, this effect was replicated here.

The results showed that the disambiguating word given the previous context is more difficult in garden-path sentences compared to control sentences. There are two possible explanations for the processing difficulty. One is that the POS sequence of a garden-path sentence is less probable than that of its control sentence. The other account is that the disambiguating word in a garden-path sentence is a lower frequency word compared to that of its control sentence.

For example, slower reading time was observed in (10a) and (11a) compared to (10b) and (11b) at the disambiguating region that is bolded.

(10) Different POS at the Disambiguating Region

a. Katie laid the dress *on the floor* **onto** (-57.80) the bed.

b. Katie laid the dress *on the floor* **after** (-55.77) her mother yelled at her.

(11) Same POS at the Disambiguating Region

a. The umpire helped the child ***on*** (-42.77) third base.

b. The umpire helped the child ***to*** (-42.23) third base.

The log probability for each disambiguating word is given at the end of each sentence. As expected, the probability at the disambiguating region in (10a) and (11a) is lower than in (10b) and (11b) respectively. The disambiguating words in (10) have different POS's; Preposition in (10a) and Conjunction (10b). This suggests that the probabilities of different POS sequences can account for different reading time at the region. In (11), however, both disambiguating words are the same POS (i.e. Preposition) and the POS sequences for both sentences are identical. Instead, "on" and "to", have different frequencies and this information is reflected in the conditional probability $P(word_i|state)$. Therefore, the slower reading time in (11b) might be attributable to the lower frequency of the disambiguating word, "to" compared to "on".

3.3 Probability Re-ranking

The probability re-ranking reported in Corley and Crocker (2000) was replicated. The tagger successfully resolved the ambiguity by reanalysis when the ambiguous word was immediately followed by the disambiguating word (e.g. Without *her* **he** was lost.). If the disambiguating word did not immediately follow the ambiguous region, (e.g. Without *her* contributions **would** be very inadequate.) the ambiguity is sometimes incorrectly resolved.

When revision occurred, probability dropped more sharply at the revision point and at the disambiguation region compared to the control sen-

tences. When the ambiguity was not correctly resolved, the probability comparison correctly modeled the comparative difficulty of the garden-path sentences

Of particular interest in this study is RR ambiguity resolution. The tagger predicted the processing difficulty of the RR ambiguity with probability re-ranking. That is, the tagger initially favors the main-verb interpretation for the ambiguous *-ed* form, and later it makes a repair when the ambiguity is resolved as a past-participle.

The RR ambiguity is often categorized as a syntactic ambiguity, but the results suggest that the ambiguity can be resolved locally and its processing difficulty can be detected by a finite state model. This suggests that we should be cautious in assuming that a structural explanation is needed for the RR ambiguity resolution, and it could be that similar cautions are in order for other ambiguities usually seen as syntactic.

4 Discussion

The current study explores Corley and Crocker's model(2000) further on the model's account of human sentence processing data seen in empirical studies. Although there have been studies on a POS tagger evaluating it as a potential cognitive module of lexical category disambiguation, there has been little work that tests it as a modeling tool of syntactically ambiguous sentence processing.

The findings here suggest that a statistical POS tagging system is more informative than Crocker and Corley demonstrated. It has a predictive power of processing delay not only for lexically ambiguous sentences but also for structurally garden-pathed sentences. This model is attractive since it is computationally simpler and requires few statistical parameters. More importantly, it is clearly defined what predictions can be and cannot be made by this model. This allows systematic testability and refutability of the model unlike some other probabilistic frameworks. Also, the model training and testing is transparent and observable, and true probability rather than transformed weights are used, all of which makes it easy to understand the mechanism of the proposed model.

Although the model we used in the current study is not a novelty, the current work largely differs from the previous study in its scope of data used and the interpretation of the model for human sentence processing. Corley and Crocker clearly state that their model is strictly limited to lexical ambiguity resolution, and their test of the model was bounded to the noun-verb ambiguity. However, the findings in the current study play out differently. The experiments conducted in this study are parallel to empirical studies with regard to the design of experimental method and the test material. The garden-path sentences used in this study are authentic, most of them are selected from the cited literature, not conveniently coined by the authors. The word-by-word probability comparison between garden-path sentences and their controls is parallel to the experimental design widely adopted in empirical studies in the form of region-by-region reading or eye-gaze time comparison. In the word-by-word probability comparison, the model is tested whether or not it correctly predicts the comparative processing difficulty at the garden-path region. Contrary to the major claim made in previous empirical studies, which is that the garden-path phenomena are either modeled by syntactic principles or by structural frequency, the findings here show that the same phenomena can be predicted without such structural information.

Therefore, the work is neither a mere extended application of Corley and Crocker's work to a broader range of data, nor does it simply confirm earlier observations that finite state machines might accurately account for psycholinguistic results to some degree. The current study provides more concrete answers to what finite state machine is relevant to what kinds of processing difficulty and to what extent.

5 Conclusion

Our studies show that, at least for the sample of test materials that we culled from the standard literature, a statistical POS tagging system can predict processing difficulty in structurally ambiguous garden-path sentences. The statistical POS tagger was surprisingly effective in modeling sentence processing data, given the locality of the probability distribution. The findings in this study provide an alternative account for the garden-path effect observed in empirical studies, specifically, that the slower processing times associated with garden-path sentences are due in part to their relatively unlikely POS sequences in comparison with those of non-garden-path sentences and in part to differences in the emission probabilities that the

tagger learns. One attractive future direction is to carry out simulations that compare the evolution of probabilities in the tagger with that in a theoretically more powerful model trained on the same data, such as an incremental statistical parser (Wang et al., 2004; Roark, 2001). In so doing we can find the places where the prediction problem faced both by the HSPM and the machines that aspire to emulate it actually warrants the greater power of structurally sensitive models, using this knowledge to mine large corpora for future experiments with human subjects.

We have not necessarily cast doubt on the hypothesis that the HSPM makes crucial use of structural information, but we have demonstrated that much of the relevant behavior can be captured in a simple model. The 'structural' regularities that we observe are reasonably well encoded into this model. For purposes of initial real-time processing it could be that the HSPM is using a similar encoding of structural regularities into convenient probabilistic or neural form. It is as yet unclear what the final form of a cognitively accurate model along these lines would be, but it is clear from our study that it is worthwhile, for the sake of clarity and explicit testability, to consider models that are simpler and more precisely specified than those assumed by dominant theories of human sentence processing.

Acknowledgments

This project was supported by the Cognitive Science Summer 2004 Research Award at the Ohio State University. We acknowledge support from NSF grant IIS 0347799.

References

T. G. Bever and B. McElree. Empty categories access their antecedents during comprehension. *Linguistic Inquiry*, 19:35–43, 1988.

L Burnard. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service, 1995.

S. Corley and M. W Crocker. *The Modular Statistical Hypothesis: Exploring Lexical Category Ambiguity*. Architectures and Mechanisms for Language Processing, M. Crocker, M. Pickering. and C. Charles (Eds.) Cambridge University Press, 2000.

F. Ferreira and J. Henderson. Use of verb information in syntactic parsing: Evidence from eye movements and word-by-word self-paced reading. *Journal of Experimental Psychology*, 16: 555–568, 1986.

L. Frazier. On comprehending sentences: Syntactic parsing strategies. *Ph.D. dissertation, University of Massachusetts*, Amherst, MA, 1978.

L. Frazier and K. Rayner. Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, 14: 178–210, 1982.

V. M. Homes, J. O'Regan, and K.G. Evensen. Eye fixation patterns during the reading of relative clause sentences. *Journal of Verbal Learning and Verbal Behavior*, 20:417–430, 1981.

J. King and M. A. Just. Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, 30:580–602, 1991.

B. MacWhinney. Basic syntactic processes. *Language acquisition; Syntax and semantics, S. Kuczaj (Ed.)*, 1:73–136, 1982.

W. M. Mak, Vonk W., and H. Schriefers. The influence of animacy on relative clause processing. *Journal of Memory and Language,*, 47:50–68, 2002.

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

B. Roark. Probabilistic top-down parsing and language modeling. *Computational Linguistics*, 27 (2):249–276, 2001.

M. J. Traxler, R. K. Morris, and R. E. Seely. Processing subject and object relative clauses: evidence from eye movements. *Journal of Memory and Language*, 47:69–90, 2002.

J. C. Trueswell. The role of lexical frequency in syntactic ambiguity resolution. *Journal of Memory and Language*, 35:556–585, 1996.

A. Viterbi. Error bounds for convolution codes and an asymptotically optimal decoding algorithm. *IEEE Transactions of Information Theory*, 13: 260–269, 1967.

W. Wang, A. Stolcke, and M. P. Harper. The use of a linguistically motivated language model in conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Montreal, Canada, 2004.