An Information-Theory-Based Feature Type Analysis for the Modelling of Statistical Parsing

SUI Zhifang [†‡], ZHAO Jun [†], Dekai WU [†]

[†] Hong Kong University of Science & Technology
Department of Computer Science
Human Language Technology Center
Clear Water Bay, Hong Kong

[‡] Peking University
Department of Computer Science & Technology
Institute of Computational Linguistics
Beijing, China

suizf@icl.pku.edu.cn, zhaojun@cs.ust.hk, dekai@cs.ust.hk

Abstract

The paper proposes an information-theory-based method for feature types analysis in probabilistic evaluation modelling for statistical parsing. The basic idea is that we use entropy and conditional entropy to measure whether a feature type grasps some of the information for syntactic structure prediction. Our experiment quantitatively analyzes several feature types' power for syntactic structure prediction and draws a series of interesting conclusions.

1 Introduction

In the field of statistical parsing, various probabilistic evaluation models have been proposed where different models use different feature types [Black, 1992] [Briscoe, 1993] [Brown, 1991] [Charniak, 1997] [Collins, 1996] [Collins, 1997] [Magerman, 1991] [Magerman, 1992] [Magerman, 1995] [Eisner, 1996]. How to evaluate the different feature types' effects for syntactic parsing? The paper proposes an information-theory-based feature types analysis model, which uses the measures of predictive information quantity, predictive information gain, predictive information redundancy and predictive information summation to quantitatively analyse the different contextual feature types' or feature types combination's predictive power for syntactic structure.

In the following, Section 2 describes the probabilistic evaluation model for syntactic trees; Section 3 proposes an information-theory-based feature type analysis model; Section 4 introduces several experimental issues; Section 5 quantitatively analyses the different contextual feature types or feature types combination in the view of information theory and draws a series of conclusion on their predictive powers for syntactic structures.

2 The probabilistic evaluation model for statistical syntactic parsing

Given a sentence, the task of statistical syntactic parsing is to assign a probability to each candidate parsing tree that conforms to the grammar and select the one with highest probability as the final analysis result. That is:

$$T_{best} = \arg\max_{T} P(T \mid S) \quad (1)$$

where S denotes the given sentence, T denotes the set of all the candidate parsing trees that conform to the grammar, $P(T|S)$ denotes the probability of parsing tree T for the given sentence S.

The task of probabilistic evaluation model in syntactic parsing is the estimation of $P(T|S)$. In the syntactic parsing model which uses rule-based grammar, the probability of a parsing tree can be defined as the probability of the derivation which generates the current parsing tree for the given sentence. That is,

$$\begin{aligned} P(T \mid S) &= P(r_1, r_2, \cdots, r_n \mid S) \\ &= \prod_{i=1}^{n} P(r_i \mid r_1, r_2, \cdots, r_{i-1}, S) \quad (2) \\ &= \prod_{i=1}^{n} P(r_i \mid h_i, S) \end{aligned}$$

Where, $r_1, r_2, \cdots, r_{i-1}$ denotes a derivation rule sequence, h_i denotes the partial parsing tree derived from $r_1, r_2, \cdots, r_{i-1}$.

In order to accurately estimate the parameters, we need to select some feature types $F_1, F_2, \cdots, F_m$, depending on which we can divide the contextual condition h_i, S for predicting rule r_i into some equivalence classes, that is, $h_i, S \xrightarrow{F_1, F_2, \cdots, F_m} [h_i, S]$, so that

$$\prod_{i=1}^{n} P(r_i \mid h_i, S) \approx \prod_{i=1}^{n} P(r_i \mid [h_i, S]) \quad (3)$$

According to the equation of (2) and (3), we have the following equation:

$$P(T \mid S) \approx \prod_{i=1}^{n} P(r_i \mid [h_i, S]) \quad (4)$$

In this way, we can get a unite expression of probabilistic evaluation model for statistical syntactic parsing. The difference among the different parsing models lies mainly in that they use different feature types or feature type combination to divide the contextual condition into equivalent classes. Our ultimate aim is to determine which combination of feature types is optimal for the probabilistic evaluation model of statistical syntactic parsing. Unfortunately, the state of knowledge in this regard is very limited. Many probabilistic evaluation models have been published inspired by one or more of these feature types [Black, 1992] [Briscoe, 1993] [Charniak, 1997] [Collins, 1996] [Collins, 1997] [Magerman, 1995] [Eisner, 1996], but discrepancies between training sets, algorithms, and hardware environments make it difficult, if not impossible, to compare the models objectively. In the paper, we propose an information-theory-based feature type analysis model by which we can quantitatively analyse the predictive power of different feature types or feature type combinations for syntactic structure in a systematic way. The conclusion is expected to provide reliable reference for feature type selection in the probabilistic evaluation modelling for statistical syntactic parsing.

3 The information-theory-based feature type analysis model for statistical syntactic parsing

In the prediction of stochastic events, entropy and conditional entropy can be used to evaluate the predictive power of different feature types. To predict a stochastic event, if the entropy of the event is much larger than its conditional entropy on condition that a feature type is known, it indicates that the feature type grasps some of the important information for the predicted event.

According to the above idea, we build the information-theory-based feature type analysis model, which is composed of four concepts: predictive information quantity, predictive information gain, predictive information redundancy and predictive information summation.

- **Predictive Information Quantity (PIQ)**

$PIQ(F; R)$, the predictive information quantity of feature type *F* to predict derivation rule *R*, is defined as the difference between the entropy of *R* and the conditional entropy of *R* on condition that the feature type *F* is known.

$$PIQ(F; R) = H(R) - H(R \mid F)$$

$$= \sum_{f \in F, r \in R} P(f, r) \log \frac{P(f, r)}{P(f) \cdot P(r)} \quad (5)$$

Predictive information quantity can be used to measure the predictive power of a feature type in feature type analysis.

- **Predictive Information Gain (PIG)**

For the prediction of rule *R*, $PIG(F_x; R|F_1, F_2, ..., F_i)$, the predictive information gain of taking F_x as a variant model on top of a baseline model employing $F_1, F_2, ..., F_i$ as feature type combination, is defined as the difference between the conditional entropy of predicting *R* based on feature type combination $F_1, F_2, ..., F_i$ and the conditional entropy of predicting *R* based on feature type combination $F_1, F_2, ..., F_i, F_x$.

$$PIG(F_x; R \mid F_1, \cdots, F_i) = H(R \mid F_1, \cdots, F_i) - H(R \mid F_1, \cdots, F_i, F_x)$$

$$= \sum_{\substack{f_1 \in F_1 \\ \cdots \\ f_i \in F_i \\ f_x \in F_x \\ r \in R}} P(f_1, \cdots, f_i, f_x, r) \log \frac{P(f_1, \cdots, f_i, f_x, r)}{P(f_1, \cdots, f_i, f_x)} \cdot \frac{P(f_1, \cdots, f_i)}{P(f_1, \cdots, f_i, r)} \quad (6)$$

If $PIG(F_x; R \mid F_1, F_2, \cdots, F_i) > PIG(F_y; R \mid F_1, F_2, \cdots, F_i)$, then F_x is deemed to be more informative than F_y for predicting *R* on top of $F_1, F_2, ..., F_i$, no matter whether $PIQ(F_x; R)$ is larger than $PIQ(F_y; R)$ or not.

- **Predictive Information Redundancy(PIR)**

Based on the above two definitions, we can further draw the definition of predictive

information redundancy as follows. $PIR(F_x,\{F_1,F_2,...,F_i\};R)$ denotes the redundant information between feature type F_x and feature type combination $\{F_1,F_2,...,F_i\}$ in predicting R, which is defined as the difference between $PIQ(F_x;R)$ and $PIG(F_x;R|F_1,F_2,...,F_i)$. That is,

$$PIR(F_X,\{F_1,F_2,\cdots,F_i\};R) = PIQ(F_X;R) - PIG(F_X;R \mid F_1,F_2,\cdots,F_i) \quad (7)$$

Predictive information redundancy can be used as a measure of the redundancy between the predictive information of a feature type and that of a feature type combination.

- **Predictive Information Summation (PIS)**

$PIS(F_1,F_2,...,F_m;R)$, the predictive information summation of feature type combination $F_1,F_2,...,F_m$, is defined as the total information that $F_1,F_2,...,F_m$ can provide for the prediction of a derivation rule. Exactly,

$$PIS(F_1,F_2,\cdots,F_m;R) = PIQ(F_1;R) + \sum_{i=2}^{m} PIG(F_i;R \mid F_1,\cdots,F_{i-1}) \quad (8)$$

4 Experimental Issues

4.1 The classification of the feature types

The predicted event of our experiment is the derivation rule to extend the current non-terminal node. The feature types for prediction can be classified into two classes, history feature types and objective feature types. In the following, we will take the parsing tree shown in Figure-1 as the example to explain the classification of the feature types.

In Figure-1, the current predicted event is the derivation rule to extend the framed non-terminal node VP, the part connected by the solid line belongs to history feature types, which is the already derived partial parsing tree, representing the structural environment of the current non-terminal node. The part framed by the larger rectangle belongs to the objective feature types, which is the word sequence containing the leaf nodes of the partial parsing tree rooted by the current node, representing the final objectives to be derived from the current node.

4.2 The corpus used in the experiment

The experimental corpus is derived from Penn TreeBank[Marcus,1993]. We semi-automatically assign a headword and a POS tag to each non-terminal node. 80% of the corpus (979,767 words) is taken as the training set, used for estimating the various co-occurrence probabilities, 10% of the corpus (133,814 words) is taken as the testing set, used to calculate predictive information quantity, predictive information gain, predictive information redundancy and predictive information summation. The other 10% of the corpus (133,814 words) is taken as the held-out set. The grammar rule set is composed of 8,126 CFG rules extracted from Penn TreeBank.

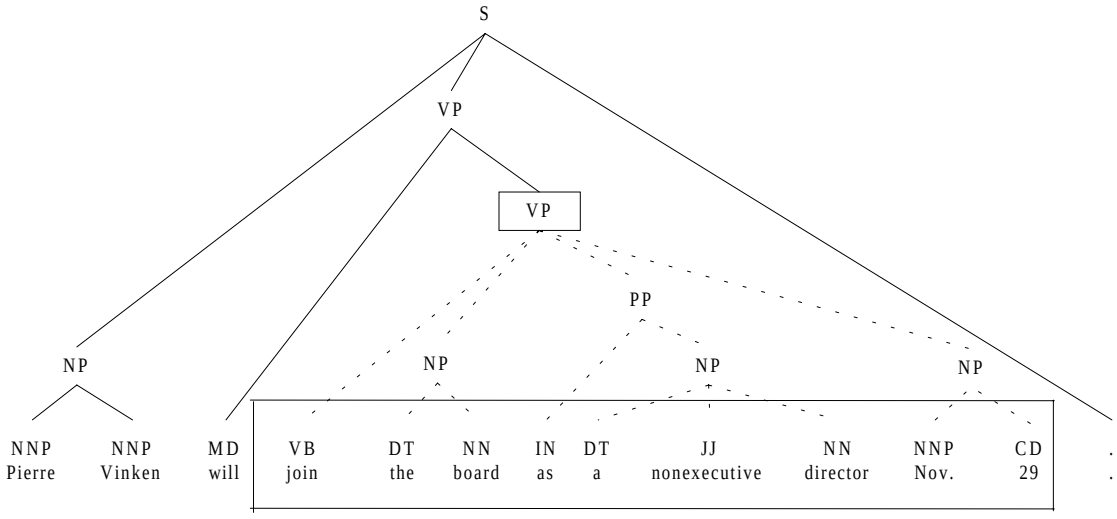

Figure-1: The classification of feature types

4.3 The smoothing method used in the experiment

In the information-theory-based feature type analysis model, we need to estimate joint probability $P(f_1, f_2, \cdots, f_i, r)$. Let F_1,F_2,...,F_i be the feature type series selected till now, $f_1 \in F_1, f_2 \in F_2, \cdots, f_i \in F_i, r \in R$, we use a blended probability $\tilde{P}(f_1, f_2, \cdots, f_i, r)$ to approximate probability $P(f_1, f_2, \cdots, f_i, r)$ in order to solve the sparse data problem[Bell, 1992].

$$\begin{aligned}&\tilde{P}(f_1, f_2, \cdots, f_i, r)\\ &= w_{-1}P_{-1}(r) + w_0 P_0(r) + \sum_{j=1}^{i} w_j P(f_1, f_2, \cdots, f_j, r)\end{aligned} \quad (9)$$

In the above formula,

$$P_{-1}(r) = \frac{1}{\sum_{\hat{r} \in R} c(\hat{r})} \quad (10)$$

$$P_0(r) = \frac{c(r)}{\sum_{\hat{r} \in R} c(\hat{r})} \quad (11)$$

where $c(r)$ is the total number of time that r has been seen in the corpus.

According to the escape mechanism in [Bell, 1992], we define the weights w_k $(-1 < k \le i)$ in the formula (9) as follows.

$$\begin{aligned}&w_k = (1 - e_k) \prod_{s=k+1}^{i} e_s \quad , \quad -1 \le k \le i\\ &w_i = 1 - e_i\end{aligned} \quad (12)$$

where e_k denotes the escape probability of context $(f_1, f_2, \cdots, f_k)$, that is, the probability in which $(f_1, f_2, \ldots, f_k, r)$ is unseen in the corpus. In such case, the blending model has to escape to the lower contexts to approximate $P(f_1, f_2, \cdots, f_k, r)$. Exactly, escape probability is defined as

$$e_k = \begin{cases} \dfrac{\sum_{\hat{r} \in R} d(f_1, f_2, ..., f_k, \hat{r})}{\sum_{\hat{r} \in R} c(f_1, f_2, ..., f_k, \hat{r})}, & 0 \le k \le i \\ 0, & k = -1 \end{cases} \quad (13)$$

where

$$d(f_1, f_2, ..., f_k, \hat{r}) = \begin{cases} 1, & if\ c(f_1, f_2, ..., f_k, \hat{r}) > 0 \\ 0, & if\ c(f_1, f_2, ..., f_k, \hat{r}) = 0 \end{cases} \quad (14)$$

In the above blending model, a special probability $P_{-1}(r) = \dfrac{1}{\sum_{\hat{r} \in R} c(\hat{r})}$ is used, where all derivation rules are given an equal probability. As a result, $\tilde{P}(f_1, f_2, \cdots, f_i, r) > 0$ as long as $\sum_{\hat{r} \in R} c(\hat{r}) > 0$.

5 The information-theory-based feature type analysis

The experiments led to a number of interesting conclusions on the predictive power of various feature types and feature type combinations, which is expected to provide reliable reference for the modelling of probabilistic parsing.

5.1 The analysis to the predictive information quantities of lexical feature types, part-of-speech feature types and constituent label feature types

- **Goal**

One of the most important variation in statistical parsing over the last few years is that statistical lexical information is incorporated into the probabilistic evaluation model. Some statistical parsing systems show that the performance is improved after the lexical information is added. Our research aims at a quantitative analysis of the differences among the predictive information quantities provided by the lexical feature types, part-of-speech feature types and constituent label feature types from the view of information theory.

- **Data**

The experiment is conducted on the history feature types of the nodes whose structural distance to the current node is within 2.

In Table-1, "Y" in *PIQ*(X of Y; R) represents the node, "X" represents the constitute label, the headword or POS of the headword of the node. In the following, the units of PIQ are bits.

- **Conclusion**

Among the feature types in the same structural position of the parsing tree, the predictive information quantity of lexical feature type is larger than that of part-of-speech feature type, and the predictive information quantity of part-of-speech feature type is larger than that of the constituent label feature type.

Table-1: The predictive information quantity of the history feature type candidates

PIQ(X of Y; R)	X= constituent label	X= headword	X= POS of the headword
Y= the current node	2.3609	3.7333	2.7708
Y= the parent	1.1598	2.3253	1.1784
Y= the grandpa	0.6483	1.6808	0.6612
Y= the first right brother of the current node	0.4730	1.1525	0.7502
Y= the first left brother of the current node	0.5832	2.1511	1.2186
Y= the second right brother of the current node	0.1066	0.5044	0.2525
Y= the second left brother of the current node	0.0949	0.6171	0.2697
Y= the first right brother of the parent	0.1068	0.3717	0.2133
Y= the first left brother of the parent	0.2505	1.5603	0.6145

5.2 The analysis to the influence of the structural relation and the structural distance to the predictive information quantities of the history feature types

- **Goal:**

In this experiment, we wish to find out the influence of the structural relation and structural distance between the current node and the node that the given feature type related to has to the predictive information quantities of these feature types.

- **Data:**

In Table-2, SR represents the structural relation between the current node and the node that the given feature type related to. SD represents the structural distance between the current node and the node that the given feature type related to.

Table-2: The predictive information quantity of the selected history feature types

PIQ(constituent label of Y; R)	SR= parent relation	SR= brother relation	SR= mixed parent and brother relation
SD=1	1.1598 (Y= the parent)	0.5832 (Y= the first left brother)	0.2505 (Y= the first left brother of the parent)
		0.4730 (Y= the first right brother)	
SD=2	0.6483 (Y= the grandpa)	0.0949 (Y= the second left brother)	0.1068 (Y= the first right brother of the parent)
		0.1066 (Y= the second right brother)	

- **Conclusion**

Among the history feature types which have the same structural relation with the current node (the relations are both parent-child relation, or both brother relation, etc), the one which has closer structural distance to the current node will provide larger predictive information quantity; Among the history feature types which have the same structural distance to the current node, the one which has parent relation with the current node will provide larger predictive information quantity than the one that has brother relation or mixed parent and brother relation to the current node (such as the parent's brother node).

5.3 The analysis to the predictive information quantities of the history feature types and the objective feature types

- **Goal**

Many of the existing probabilistic evaluation models prefer to use history feature types other than objective feature types. We select some of history feature types and objective feature types, and quantitatively compare their predictive information quantities.

- **Data**

The history feature type we use here is the headword of the parent, which has the largest predictive information quantity among all the history feature types. The objective feature types are selected stochastically, which are the first

word and the second word in the objective word sequence of the current node (Please see 4.1 and Figure-1 for detailed descriptions on the selected feature types).

Table-3: The predictive information quantity of the selected history and objective feature types

Class	Feature type	PIQ(Y;R)
History feature type	Y= headword of the parent	2.3253
Objective feature type	Y= the first word in the objective word sequence	3.2398
	Y= the second word in the objective word sequence	3.0071

- **Conclusion**

Either of the predictive information quantity of the first word and the second word in the objective word sequence is larger than that of the headword of the parent node which has the largest predictive information quantity among all of the history feature type candidates. That is to say, objective feature types may have larger predictive power than that of the history feature type.

5.4 The analysis to the predictive information quantities of the objective features types selected respectively on the physical position information, the heuristic information of headword and modifier, and the exact headword information

- **Goal**

Not alike the structural history feature types, the objective feature types are sequential. Generally, the candidates of the objective feature types are selected according to the physical position. However, from the linguistic viewpoint, the physical position information can hardly grasp the relations between the linguistic structures. Therefore, besides the physical position information, our research try to select the objective feature types respectively according to the exact headword information and the heuristic information of headword and modifier. Through the experiment, we hope to find out what influence the exact headword information, the heuristic information of headword and modifier, and the physical position information have respectively to the predictive information quantities of the feature types.

- **Data:**

Table-4 gives the evidence for the claim.

Table-4: the predictive information quantity of the selected objective feature types

the information used to select the objective feature types	PIQ(Y;R)
the physical position information	3.2398 (Y= the first word in the objective word sequence)
Heuristic information 1: determine whether a word has the possibility to act as the headword of the current constitute according to its POS	3.1401 (Y= the first word in the objective word sequence which has the possibility to act as the headword of the current constitute)
Heuristic information 2: determine whether a word has the possibility to act as the modifier of the current constitute according to its POS	3.1374 (Y= the first word in the objective word sequence which has the possibility to act as the modifier of the current constitute)
Heuristic information 3: given the current headword, determine whether a word has the possibility to modify the headword	2.8757 (Y= the first word in the objective word sequence which has the possibility to modify the headword)
the exact headword information	3.7333 (Y= the headword of the current constitute)

- **Conclusion**

The predictive information quantity of the headword of the current node is larger than that of a feature type selected according to the selected heuristic information of headword or modifier, and larger than that of a feature type selected according to the physical positions; The

predictive information quantity of a feature type selected according to the physical positions is larger than that of a feature types selected according to the selected heuristic information of headword or modifier.

5.5 The selection of the feature type combination which has the optimal predictive information summation

- **Goal:**

We aim at proposing a method to select the feature types combination that has the optimal predictive information summation for prediction.

- **Approach**

We use the following greedy algorithm to select the optimal feature type combination.

In building a model, the first feature type to be selected is the feature type which has the largest predictive information quantity for the prediction of the derivation rule among all of the feature type candidates, that is,

$$F_1 = \arg\max_{F_i \in \Omega} PIQ(F_i; R) \qquad (15)$$

Where Ω is the set of candidate feature types.

Given that the model has selected feature type combination $F_1, F_2, \cdots, F_j$, the next feature type to be added into the model is the feature type which has the largest predictive information gain in all of the feature type candidate except $F_1, F_2, \cdots, F_j$, on condition that $F_1, F_2, \cdots, F_j$ is known. That is,

$$F_{j+1} = \arg\max_{\substack{F_i \in \Omega \\ F_i \notin \{F_1, F_2, \cdots, F_j\}}} PIG(F_i; R \mid F_1, F_2, \cdots, F_j) \qquad (16)$$

- **Data:**

Among the feature types mentioned above, the optimal feature type combination (i.e. the feature type combination with the largest predictive information summation) which is composed of 6 feature types is, the headword of the current node (type1), the headword of the parent node (type2), the headword of the grandpa node (type3), the first word in the objective word sequence(type4), the first word in the objective word sequence which have the possibility to act as the headword of the current constitute(type5), the headword of the right brother node(type6). The cumulative predictive information summation is showed in Figure-2

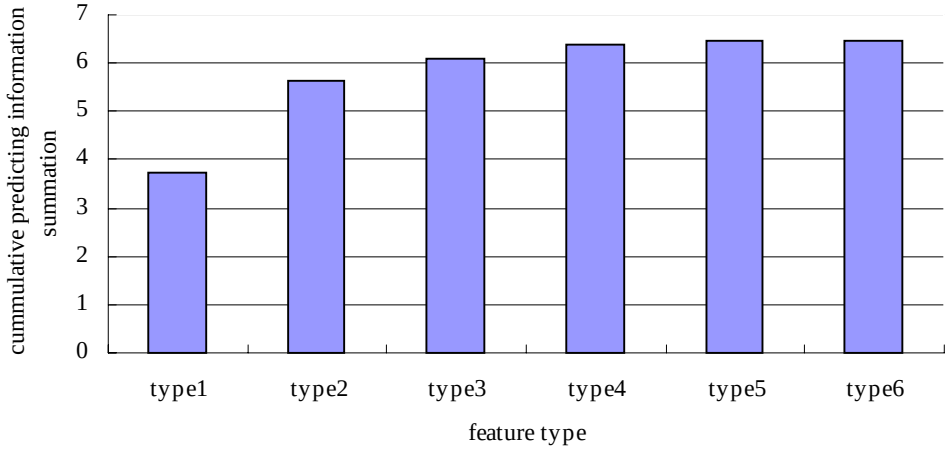

Figure-2: The cumulative predictive information summation of the feature type combinations

6 Conclusion

The paper proposes an information-theory-based feature type analysis method, which not only presents a series of heuristic conclusion on the predictive power of the different feature types and feature type combination for syntactic parsing, but also provides a guide for the modeling of syntactic parsing in the view of methodology, that is, we can quantitatively analyse the different contextual feature types or feature types combination's effect for syntactic

structure prediction in advance. Based on these analysis, we can select the feature type or feature types combination that has the optimal predictive information summation to build the probabilistic parsing model.

However, there are still some questions to be answered in this paper. For example, what is the beneficial improvement in the performance after using this method in a real parser? Whether the improvements in PIQ will lead to the improvement of parsing accuracy or not? In the following research, we will incorporate these conclusions into a real parser to see whether the parsing accuracy can be improved or not. Another work we will do is to do some experimental analysis to find the impact of data sparseness on feature type analysis, which is critical to the performance of real systems.

The proposed feature type analysis method can be used in not only the probabilistic modelling for statistical syntactic parsing, but also language modelling in more general fields [WU, 1999a] [WU, 1999b].

References

Bell, T.C., Cleary, J.G., Witten,I.H. 1992. Text Compression, PRENTICE HALL, Englewood Cliffs, New Jersey 07632, 1992

Black, E., Jelinek, F.,Lafferty, J.,Magerman, D.M., Mercer, R. and Roukos, S. 1992. Towards history-based grammars: using richer models of context in probabilistic parsing. In Proceedings of the February 1992 DARPA Speech and Natural Language Workshop, Arden House, NY.

Brown, P., Jelinek, F., & Mercer, R. 1991. Basic method of probabilistic context-free grammars. IBM internal Report, Yorktown Heights, NY.

T.Briscoe and J. Carroll. 1993. Generalized LR parsing of natural language (corpora) with unification-based grammars. Computational Linguistics, 19(1): 25-60

Eugene Charniak. 1997. Statistical parsing with a context-free grammar and word statics. In Proceedings of the Fourteenth National Conference on Artificial Intelligence, AAAI Press/MIT Press, Menlo Park.

Stanley F. Chen and Joshua Goodman. 1999. An Empirical Study of Smoothing Techniques for Language Modeling. Computer Speech and Language, Vol.13, 1999

Michael John Collins. 1996. A new statistical parser based on bigram lexical dependencies. In Proceedings of the 34th Annual Meeting of the ACL.

Michael John Collins. 1997. Three generative lexicalised models for statistical parsing. In Proceedings of the 35th Annual Meeting of the ACL.

J.Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In Proceedings of COLING-96, pages 340-345

Joshua Goodman. 1998. Parsing Inside-Out. PhD. Thesis, Harvard University, 1998

Magerman, D.M. and Marcus, M.P. 1991. Pearl: a probabilistic chart parser. In Proceedings of the European ACL Conference, Berlin, Germany.

Magerman, D.M. and Weir, C. 1992. Probabilistic prediction and Picky chart parsing. In Proceedings of the February 1992 DARPA Speech and Natural Language Workshop, Arden House, NY.

David M. Magerman. 1995. Statistical decision-tree models for parsing. In Proceedings of the 33th Annual Meeting of the ACL.

Mitchell P. Marcus, Beatrice Santorini & Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. Computational Linguistics 19, pages 313-330

C. E. Shannon. 1951. Prediction and Entropy of Printed English. Bell System Technical Journal, 1951

Dekai,Wu, Sui Zhifang, Zhao Jun. 1999a. An Information-Based Method for Selecting Feature Types for Word Prediction. Proceedings of Eurospeech'99, Budapest Hungary

Dekai, Wu, Zhao Jun, Sui Zhifang. 1999b. An Information-Theoretic Empirical Analysis of Dependency-Based Feature Types for Word Prediction Models. Proceedings of EMNLP'99, University of Maryland, USA