A Proposal of the Partition on Vocabulary

Mutsumi Sato and Kohkichi Tanaka**

Abstract

In this paper an analytical model of the language is considered. The "derivative" of the partition R on a vocabulary and the "matched partition" of the language are defined and some properties of them are considered. The "derivative" defined in this paper does not mean the "derivative" in Marcus[1]. And the "matched" means that a word "a" is equal to a word "b" in the sense of /R if and only if the context of "a" is equal to that of "b" in the sense of /R.

We propose the partition P: the maximum member of the class of the matched partition.

1. Introduction

Some analytical models of languages have been studied.[1),2)] In their models the distributional partition has been used for analyzing languages.

In this paper we introduce the derivative of the partition and the matched partition. The distributinal partition is the minimum member of matched partitions. We propose the maximum matched partition instead of the distributional partition.

2. Definitions and Fundamental Theorems

2.1 Definitions

In this section we introduce some notations and definitions.

Definition 2.1 Σ is a finite set called vocabulary. The elements of Σ, denoted by a, b, c,..., are called words. Σ^* is the set of all finite strings of words. The elements of Σ^*, denoted by α, β, γ,..., are called sentences. The zero string, denoted by ϵ, is a string such that $\epsilon\alpha = \alpha\epsilon = \alpha$ for each string α.

Definition 2.2 A subset L of Σ^* is a language over Σ.

This paper first appeared in Japanese in Joho-Shori (Journal of the Information Processing Society of Japan), Vol. 16, No. 2 (1975), pp. 102~107.

* Dept. of IE., Faculty of Science & Technology, Kinki University
** Dept. of Information Science, Faculty of Engineering Science, Osaka University

Definition 2.3 Partitions of Σ are denoted by R, R',... . Each set of partition R will be called a cell of R. A partition R is denoted as $R=\bigcup_{i=1}^{n} R_i$; the union of all cells. We identify a equivalence relation R on Σ with a partition R of Σ and equivalence classes with cells.

Definition 2.4 If $R(a) \subseteq R'(a)$ for each word a, then $R \leq R'$(R is finer than R'). The union of R and R' is denoted by $R \vee R'$ and the intersection by $R \wedge R'$.

Definition 2.5 A partition in which each cell is formed by a single word is called the unit partition of Σ and denoted by E. A partition which has a single cell is called the improper partition of Σ and denoted by Σ.

Definition 2.6 A homomorphism naturally induced by the equivalence relation R is denoted by /R. The homomorphism /R is extended to Σ^*, $\Sigma^* \times \Sigma^*$, 2^{Σ^*}, $2^{\Sigma^* \times \Sigma^*}$.

i.e. $a/R=R(a)$

$\alpha \cdot \beta/R=\alpha/R \cdot \beta/R$

$(\alpha, \beta)/R=(\alpha/R, \beta/R)$

$\{\alpha\}/R=\{\alpha/R\}$

$\{(\alpha,\beta)\}/R=\{(\alpha,\beta)/R\}$

Definition 2.7 The set of contexts of a word a is denoted by $C_L(a)$.

i.e. $C_L(a)=\{(\alpha,\beta) \mid \alpha a \beta \in L\}$

Definition 2.8 The partition $\sim_L/R$ defined as follows is called the derivative of the partition R.

$$a \sim_L/R\ b \quad \text{iff} \quad C_L(a)/R=C_L(b)/R.$$

The derivative defined above does not mean the "derivative" in Marcus[1]. In Marcus the "derivative" of the partition R is not finer than R, but in this paper there is a partition R more rough than the derivative of itself.

Henceforth, the subscript L on the $C_L(a)$ and $\sim_L/R$ is dropped whenever L is clearly understood.

Definition 2.9 For a given language L, R which is a partition with the property

(1) $R=\sim/R$

is called a matched partition of L.

Let R is a matched partition. If two words a and b are contained in a same cell (aRb), then a and b have same context in the sense of /R(i.e. C(a)/R=C(b)/R).

Example Let

(vocabulary) $\Sigma=\{$I, You, have, pens, desks$\}$,

(language) $L=\{$I have pens, You have desks$\}$,

(partition) $R=\{I, You\}\cup\{have\}\cup\{pens, desks\}$,

then $\sim/\Sigma=R$, $\sim/R=R$: i.e. R is a matched partition.

2.2 Fundamental theorems

In this sectition fundamental theorems on derivatives are described.

Theorem 2.1 Let R is finer than R'. If $C(a)/R\subseteq C(b)/R$, then $C(a)/R'\subseteq C(b)/R'$.

In view of theorem 2.1 following two corollaries are easily shown.

Corollary 2.1.1 Let R is finer than R'. If $C(a)/R=C(b)/R$, then $C(a)/R'=C(b)/R'$.

Corollary 2.1.2 If R is finer than R' then $\sim/R$ is also finer than $\sim/R'$(the derivative of R is finer than that of R').

The inverse of corollary 2.1.2 is not hold.

Example Let $\Sigma=\{a, b, c\}$, $L=\{ab, ba, bc, cb, ca, ac, cc, aa\}$, $R=\{a, c\}\cup\{b\}$, $R'=\{a, b\}\cup\{c\}$, then $\sim/R'=\{a, b, c\}\gtrsim\sim/R=R$ and $R\lneq R'$.

Theorem 2.2 Let $\{R_i\}=\{R_1, R_2, \ldots, R_n\}$, then

(i) $\bigwedge_{i=1}^{n}\sim/R_i \geq \sim/(\bigwedge_{i=1}^{n}R_i)$

(ii) $(\bigvee_{i=1}^{n}\sim/R_i)^* \leq \sim/(\bigvee_{i=1}^{n}R_i)^*$

where * means reflexive and transitive closure of relations.

3. Derivative sequences

In this section some properties of derivative sequences are described.

Definition 3.1 A derivative sequence ${}^iR$ is defined as follows:

${}^iR={}_0R, {}_1R, {}_2R, \ldots$, where ${}_0R=R$, ${}_iR=/_{(i-1)}R$.

Definition 3.2 A derivative sequence ${}^iR$ or initial partition ${}_0R$ is said to be converged if there exists a natural number N, for any $n(n\geq N)$, ${}_NR={}_nR$. The converged partition ${}_NR$ is denoted by ${}_\infty R$.

Obviously the converged partition ${}_\infty R$ is a matched partition.

Theorem 3.1 If ${}_0R$ is finer(more rough) than ${}_1R$ then ${}_iR$ is finer(more rough) than ${}_{i+1}R$ for nonnegative integer i.

Theorem 3.2 If ${}_0R$ is finer(more rough) than ${}_1R$ then the derivative sequence ${}^iR$ converges.

Theorem 3.3 Assume that R and R' converge. If R is finer than R' then ${}_\infty R$ is finer than ${}_\infty R'$.

Corollary 3.3.1 Σ and E always converge and ${}_\infty\Sigma\geq{}_\infty E$.

Theorem 3.4 $\sim/E={}_\infty E$, but $\sim/\Sigma$ is not always equal to ${}_\infty\Sigma$.

Theorem 3.5 There exists a derivative sequence which does not converge.

Example Let $\Sigma=\{c, d, e, f\}$, $L=\{ce, df\}$, $R=\{c, d\}\vee\{e\}\vee\{f\}$, then ${}^{1}R$ does not converge. Because ${}_{2i}R=R$, ${}_{2i+1}R=\{c\}\vee\{d\}\vee\{e, f\}$ for any natural number i.

Generally a derivative sequence converges or fall into ultimately cyclic as above example.

4. Matched partition

In this section it is shown that the set of matched partitions forms a lattice and the maximum member in it is ${}_{\infty}\Sigma$ and the minimum member ${}_{\infty}E$.

Definition 4.1 ${}_{\infty}\Sigma$, denoted by P, is called the maximum matched partition and ${}_{\infty}E$, denoted by $\sim$, is called the minimum matched partition(or the distributional partition).

Lemma 4.1.1 Let R is a matched partition. $P\geq R\geq\sim$.

Lemma 4.1.2 Let R and R' are matched partitions. $R\wedge R'$ converges and ${}_{\infty}(R\wedge R')\leq R\wedge R'$.

Lemma 4.1.3 Let R and R' are matched partitions. $(R\vee R')^{*}$ converges and ${}_{\infty}(R\vee R')^{*}\geq (R\vee R')^{*}$.

There exists two matched partitions such that $R\wedge R'$(or $(R\vee R')^{*}$) is not a matched partition.

Theorem 4.1 The set of matched partitions forms a lattice.

5. Proposal of the maximum matched partition P

In this section the reason why we propose the maximum matched partition P is explained.

A partition used in analysis of syntax will be requested the following properties.

(i) Matched partition.

(ii) Independent of semantics

(iii) Time-stability.

(iv) Simpleness of grammar.

We propose the maximum matched partition P by the following reasons.

(i) P is clearly a matched partition.

(ii) P is independent of semantics because P is obtained from Σ through repetition of derivations.

(iii) P is, therefore, stable in time.

(iv) P simplify the grammar because there exists a cotext sensitive language L such

that L/P is a context free language.

6. Summary

We introduce the derivative and the matched partition and show some properties of them. And we propose the maximum matched partition P, intending to analyze languages.

The above theory, however, may not be immediately applied to the analysis of natural languages(c.f. see the following appendix).

The authors wish to express their thanks to Dr. Tadahiro Kitahashi of Osaka University and Prof. Iichi Taki of Kinki University.

References

1) S. Marcus : Algebraic Linguistics; Analytical Models, Academic Press, New York & London(1967).

2) F. Kiefer: Mathematical Linguistics in Eastern Europe, Elsevier, New York(1968).

Appendix

The λ-derivative and the λ-matched partition act as substitute for the derivative and the matched partition respectively in the above theory.

The λ-derivative of R and the λ-matched partition will be defined as follows.

Let the language L is finite.

The number of elements in a finite set S will be denoted by n(S).

$C_b^R(a) \triangleq \{(\alpha, \beta) \mid (\alpha, \beta) \in C(a),\ (\alpha, \beta)/R \in C(b)/R\}$

$$F_R(a, b) \triangleq \frac{n(C_b^R(a)) + n(C_a^R(b))}{n(C(a)) + n(C(b))}$$

$a \cdot A_\lambda(R) \cdot b$ iff $F_R(a, b) \geq \lambda$, where $0 < \lambda \leq 1$.

The reflexive and transitive closure $A_\lambda^*(R)$ of $A_\lambda(R)$ is said to be the λ-derivative of R.

The partition R is said to be the λ-matched partition if $R = A_\lambda^*(R)$.