A Model of Random Merging of Segments

L. G. Mityushin

Kharkevich Institute for Information Transmission Problems,
Russian Academy of Sciences, Moscow
mit@iitp.ru

Received November 13, 2012; in final form, February 6, 2013

Abstract—We consider a growing set $U$ of segments with integer endpoints on a line. For every pair of adjacent segments, their union is added to $U$ with probability $q$. At the beginning, $U$ contains all segments of length from 1 to $m$. Let $h_n$ be the probability that the segment $[a, a+n]$ will be created; the critical value $q_c(m)$ is defined as $\sup \{q \mid \lim_{n \to \infty} h_n = 0\}$. Lower and upper bounds for $q_c(m)$ are obtained.

DOI: 10.1134/S003294601303006X

1. INTRODUCTION

We consider segments with integer endpoints on a line which “enter into reaction” with each other. Any two adjacent segments can react (adjacent means having a single common point). As a result of a reaction, a new segment, which is the union of the original ones, is created with probability $q$; the original segments are retained. The probability $q$ is the same for all reactions, and the results of reactions are mutually independent. Equal segments produced by different reactions are considered to be the same. Every pair of adjacent segments reacts once; new segments react both between themselves and with segments that existed before.

This model arose in connection with natural language parsing. Some text segments (parts of sentences or whole sentences) can be assigned syntactic structures which represent grammatical relations between words and/or groups of words in a segment. The two types of syntactic structures used most widely are phrase structure trees and dependency trees (see, e.g., [1,2]). In these trees, vertices are assigned labels of grammatical categories (and possibly some others), and arcs may be assigned labels of syntactic relations. Syntactic structures must satisfy certain requirements called “grammar rules.” The aim of parsing is to find the syntactic structure or structures which correspond to a given text segment. It should be noted that real syntactic structures are typically nonunique: if, on a certain text segment of length $n$, at least one structure can be built, then the number of different structures that can be built on this segment is “on average” of order $\exp(Cn)$.

An important property of syntactic structures is the possibility of constructing them recursively. Assume that at first we have structures for segments of length 1, i.e., separate words. Then structures built on adjacent segments are used to build structures on unions of those segments. In the cases of phrase structure trees and dependency trees, the tree that represents the structure on the union of segments contains the original structures as subtrees, and also contains additional elements that unite them into a single tree.

Paper [3] describes an approach to parsing based on the idea of local interaction. Linguistic quality of syntactic structures is assumed to be measured by means of a certain objective function. There is a growing set $U$ of syntactic structures which at the initial moment contains structures on segments of length 1. Then pairs of structures on adjacent segments interact “like molecules in a test tube,” and if, for a given pair of structures, a new structure can be built on the union of their
segments, this structure is added to $U$ with a certain probability $q$. The value of $q$ is determined by an external control mechanism, which gives higher values of $q$ to structures with higher values of the objective function. As a result, for a given sentence, its optimal or suboptimal syntactic structures are found fairly quickly.

The model that we consider is a maximal simplification of this situation. It is not supposed to have any linguistic significance and is studied as a purely mathematical object. Some natural generalizations are possible: for example, the probability $q$ of generating a new segment can depend on its length $n$; this dependence may be of the form $q = \min(1, cn^{-d})$, where $c > 0$ and $0 < d \leq 1$ are parameters. Another option is a model with many types of segments in which segments of types depend on the length of the new segment.

Let us return to the original model. It is convenient to describe the creation of segments as a discrete-time process in which segments of length $n$ are created at moment $n$. Let $s_n^a$ be the indicator of the event “the system contains a segment of length $n$ with the left endpoint $a$.” Assume that at the beginning the system contains all the segments of length from 1 to $m$, and them only. Then $s_n^a = 1$ for $n \leq m, a \in \mathbb{Z}$; the values of $s_n^a, n > m$, are determined as follows. Let $\xi_n^{ab}, n > 1, a, b \in \mathbb{Z}, a < b \leq a + n$, be mutually independent random variables that are equal to 1 with probability $q$ and to 0 with probability $1 - q$. The variable $\xi_n^{ab}$ determines the result of the reaction between segments $[a, b]$ and $[b, a + n]$. The values $s_n^a$ are found successively for $n = m + 1, m + 2 \ldots$ as follows:

$$s_n^a = \max_{0 < k < n} s_k^a s_{n-k}^{a+k},$$

which reflects the $n - 1$ different ways of building a segment of length $n$ out of two adjacent segments.

For a fixed $m$, let $h_n = P(s_n^a = 1)$ (it does not depend on $a$). What we are interested in is asymptotic behavior of $h_n$ for various values of $q$. It is easy to show that $h_n$ is nondecreasing as a function of $q$. For a given $m$, define the critical value $q_c(m) = \sup\{q \mid \lim_{n \to \infty} h_n = 0}\}.

The following theorems provide lower and upper estimates for $q_c(m)$.

**Theorem 1.** We have $q_c(1) \geq \frac{1}{4}$ and

$$q_c(m) \geq \left(1 - \sqrt{(m+1)/2m}\right)/(m-1), \quad m > 1.$$

**Theorem 2.** We have $q_c(m) \leq 1 - 3^{-1/r} < \frac{\ln 3}{r}$, where $r = [(m+1)/2]$.

Monte Carlo simulation shows that for large $n$, the dependence of $h_n$ on $q$ is close to a step-function with values 0 and 1. The table presents the data for $m = 1$; $\overline{h_{1000}}$ is the average value of $s_1^{1000}$ obtained in 100 trials.

<table>
<thead>
<tr>
<th>$q$</th>
<th>$\overline{h_{1000}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq 0.34$</td>
<td>0.00</td>
</tr>
<tr>
<td>0.35</td>
<td>0.10</td>
</tr>
<tr>
<td>0.36</td>
<td>0.59</td>
</tr>
<tr>
<td>0.37</td>
<td>0.98</td>
</tr>
<tr>
<td>$\geq 0.38$</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Conjecture.** For any $m$ and $q > q_c(m)$, we have $\lim_{n \to \infty} h_n = 1$.

A weaker result can be proved using methods similar to the proof of Theorem 2. As above, $r = [(m+1)/2]$. 

PROBLEMS OF INFORMATION TRANSMISSION Vol. 49 No. 3 2013