Information Theoretic Analysis of Postal Address Fields for Automatic Address Interpretation

Sargur N. Srihari, Wen-jann Yang, and Venugopal Govindaraju
CEDAR, SUNY at Buffalo
Amherst, NY 14228-2583, USA
{srihari, wyang, govind}@cedar.buffalo.edu

Abstract

This paper concerns a study of information content in postal address fields for automatic address interpretation. Information provided by a combination of address components and information interaction among components is characterized in terms of Shannon's entropy. The efficiency of assignment strategies for determining a delivery point code can be compared by the propagation of uncertainty in address components. The quantity of redundancy between components can be computed from the information provided by these components. This information is useful in developing a strategy for selecting a useful component for recovering the value of an uncertain component. The uncertainty of a component based on another known component can be measured by conditional entropy. By ranking the uncertainty quantity, the effective processing flow for determining the value of a candidate component can be constructed.

1. Introduction

A tremendous number of mail pieces are handled every day by post offices in the world. In fiscal year 1997, the United States Postal Service (USPS) handled about 630 million pieces of mail a day, six days a week, for a total of 191 billion pieces. The national delivery network in the US now reaches nearly 130 million addresses [1]. To assist in processing this large volume of mail, automatic address interpretation (AI) is vital.

Most of the previous work on AI emphasizes overcoming word recognition problems [2, 3]. The street name lexicon from a postal address directory for helping handwritten word recognition has been proven to be useful [4]. Statistical information from postal addresses is useful for address interpretation beyond the task of word recognition. However, statistical analysis of postal address fields and the study of applying results of the statistical analysis to address interpretation has not been done yet.

2. Overview of automatic address interpretation

The address interpretation model is shown in Fig. 1. An address interpretation engine (*AI*) takes an image (x) as input, interprets the image content of several address fields in a destination address block, and refers to some knowledge sources ($\mathcal{S}$) for useful information in order to assign the most cost effective delivery point code (*I(x)*). The knowledge source could consist of mail stream history and a directory of legal addresses.

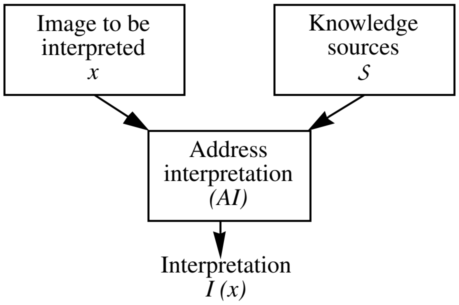
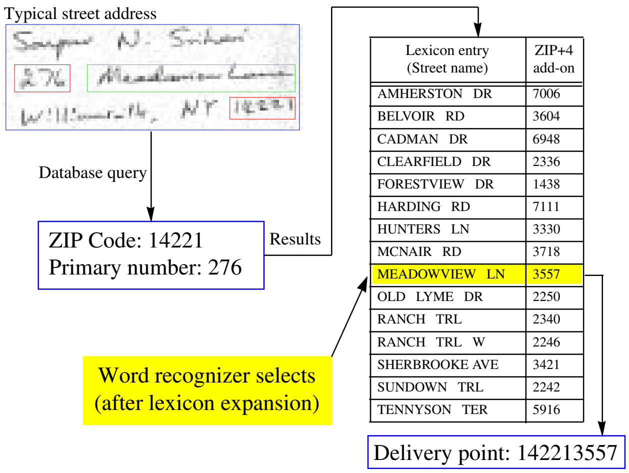


Figure 1. Address interpretation model

The goal of automatic address interpretation is to construct the most cost effective delivery point code for a destination address. An assignment strategy for determining a delivery point code is shown in Fig. 2 for a destination address. After the destination address block is located, the ZIP Code (14221) is interpreted, then the primary number (276) is interpreted, and then the street name lexicon is generated for the (ZIP Code, primary number) pair by querying an address directory. The street name lexicon is expanded to include variant names (e.g., LN vs. LANE). The word recognizer is called to get the best match street name in the lexicon (i.e., MEADONVIEW LN). Finally, the corresponding ZIP+4 add-on (i.e., 3557) is retrieved to construct the delivery point code (142213557).



Lexicon entry (Street name)	ZIP+4 add-on
AMHERSTON DR	7006
BELVOIR RD	3604
CADMAN DR	6948
CLEARFIELD DR	2336
FORESTVIEW DR	1438
HARDING RD	7111
HUNTERS LN	3330
MCNAIR RD	3718
MEADOWVIEW LN	3557
OLD LYME DR	2250
RANCH TRL	2340
RANCH TRL W	2246
SHERBROOKE AVE	3421
SUNDOWN TRL	2242
TENNYSON TER	5916

Figure 2. An assignment strategy used in CEDAR's HWAI system

3. Information measure

3.1. Measure of information

The concept of Shannon's entropy [5] for discrete source of information is used as the measure of information. *Definition:* A *component c* is an address field f_i, a portion of f_i (e.g., a digit), or a combination of components.

Three kinds of information are measured.

- The information provided (*IP*) by a component x = $\mathcal{H}(x)$ (Entropy assuming uniform distribution).

$$\mathcal{H}(x) = log_2|x| \quad (1)$$

- The uncertainty of a component y when a component x is known = $H_x(y)$ (i.e., conditional entropy). The quantity of $H_x(y)$ measures how uncertain we are of y on the average of known x. The $H_x(y)$ is defined as the average of the entropy of y for each value of x, weighted according to the probability of getting that particular x.

$$H_x(y) = -\sum_{i,j} p_{ij}(log_2 \frac{p_{ij}}{\sum_j p_{ij}}) \quad (2)$$

Where x_i is a value of the component x ;
y_j is a value of the component y ;
p_{ij} is the joint probability of $p(x_i, y_j)$.

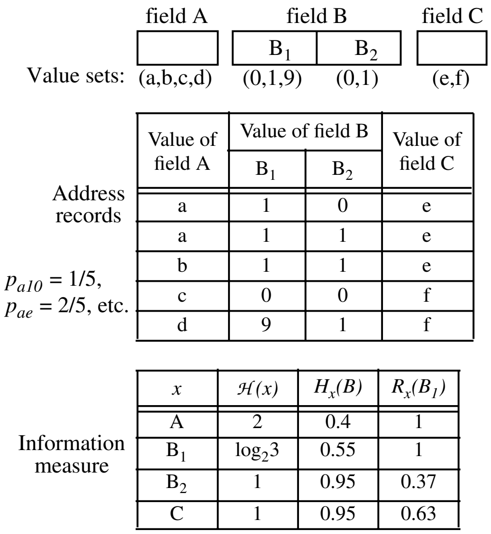
- Redundancy of components x to y is defined as follows.

$$R_x(y) = \mathcal{I}(x,y)/\mathcal{H}(y), \quad (3)$$

where $\mathcal{I}(x,y) = \mathcal{H}(x) + \mathcal{H}(y) - \mathcal{H}(x,y)$. In the computation of $\mathcal{H}(x,y)$, the fields x and y are considered as a single component. The redundancy quantity lies between 0 and 1. The larger quantity indicates that more information in the y component is shared by the x.

3.2. Example of information measure

As an illustration, suppose there are 3 fields (*A, B, C*) of interest in address interpretation and field *B* contains subfields B_1 and B_2. The possible values for each field are shown in Fig. 3. From the information measure defined in the section 3.1, the result is shown in Fig. 3. For example, there are 4 different values in the field *A*, therefore $\mathcal{H}(A) = 2$. It means that knowing the field *A* provides 2 bits of information. Since $p_{a10} = p_{a11} = p_{b11} = p_{c00} = p_{d91} = 1/5$, and $p_a = 2/5$ and $p_b = p_c = p_d = 1/5$, the uncertainty of field *B* by knowing the field *A* is computed by $H_A(B) = 0.4$ bit. Since $\mathcal{H}(A) = 2, \mathcal{H}(B_1) = log_2 3, \mathcal{H}(A, B_1) = 2$, the redundancy of field *A* to field *B* is $R_A(B_1) = 1$. The field B_1 is a completely redundant component to the field *A*. It means that if the field B_1 cannot be interpreted, the interpretation of the field *A* can recover the value in the field B_1. In this case, the field *A* is more valuable to be interpreted than the fields B_2 and *C*.



Value of field A	Value of field B		Value of field C
	B_1	B_2	
a	1	0	e
a	1	1	e
b	1	1	e
c	0	0	f
d	9	1	f

x	$\mathcal{H}(x)$	$H_x(B)$	$R_x(B_1)$
A	2	0.4	1
B_1	$log_2 3$	0.55	1
B_2	1	0.95	0.37
C	1	0.95	0.63

Figure 3. Example of information measure

4. Measure of information from a US postal address directory

4.1. Fields of interest

Two kinds of US postal address directories, National City State File and Delivery Point Files, are analyzed. The city name (f_1), state abbreviation (f_2), and ZIP Code (f_3) of every record in a National City State File are extracted and the duplicated data are eliminated. The ZIP Code (f_3), ZIP+4 add-on (f_4), primary number (f_5), street name (f_6), secondary designator abbreviation (f_7), secondary number (f_8), and building/firm name (f_9) of every record in Delivery Point Files are also extracted, and the duplicated data are eliminated. The resulting data are used for the measure of information. Any combination of f_1, f_2, and any combination of digits (f_{3i}) in f_3, is used to measure the information provided and conditional entropy for the f_3 field. Any combination of fields from f_3 to f_9 is also used to measure the information provided and conditional entropy for the f_4 field. In the measure of the conditional entropy, the probability value is from the occurrence of records of the fields of interest.

4.2. Information content in US postal address fields

- The information provided by each field is shown in Table 1. How much information is provided on the average for a field of an address can be found. For example, knowing f_1 provides 15.28 bits of information.

Table 1. Information provided by each field

Component	x	$\mathcal{H}(x)$
City name	f_1	15.28
State abbr.	f_2	5.95
ZIP Code	f_3	15.39
ZIP+4 Add-on	f_4	13.29
Primary number	f_5	20.14
Street name	f_6	20.22
Sec. designator abbr.	f_7	4.58
Sec. number	f_8	16.92
Building/firm name	f_9	19.85

- The assignment strategy for determining the delivery point code by our handwritten address interpretation(HWAI) system is shown in Fig. 2 [2]. The propagation of uncertainty of applying this strategy is shown in Fig. 4. In the condition of only knowing valid values of ZIP+4 add-on (f_4), 13.29 bits of information are provided. After a ZIP Code is known, the uncertainty for f_4 is reduced to 10.50 bits. When more information is known, less uncertainty for $f4$ is obtained.

  The Fig. 4 also shows the propagation of uncertainty of another assignment strategy which is (ZIP Code -> secondary number -> street name). The rate of uncertainty reduction is less than that shown in Fig. 4. Since it is possible that a ZIP+4 add-on can be determined by knowing the value of one or two fields, it indicates that this assignment strategy is less efficient and more fields of information may need to be identified in order to construct a ZIP+4 add-on value.

4.3. Ranking processing priority for confirming ZIP Code

Assume that the components to be recognized for determining f_3 are f_{3i}, f_2, and f_1. Based on the available information at that time, the component to be recognized next, which is the most effective, can be determined from the measure of information. For the example of determining f_3 in the condition of known f_1, the most effective component (on the average condition) to be recognized next is f_{35} (Fig. 5). The value of f_3 (ZIP-Code) uncertainty is 0.63 bit which possesses the smallest value in the column of knowing 2 components. Since the uncertainty is low, it indicates that the f_3 value may be determined at that time (average number of f_3 candidates is 1.29 when f_1 and f_{35} are known), and if it is determined, no further processing is required. If f_3 is still not determined, following the same rule, the next most effective component is f_{34}, then f_{33}, and finally f_2. The scheme of ranking processing priority can be applied for confirming other fields of interest.

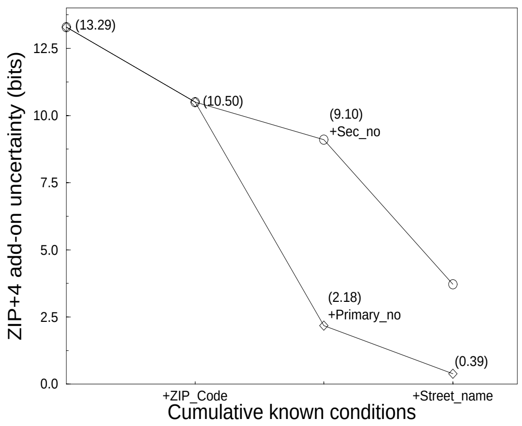


Figure 4. Propagation of uncertainty for assignment strategies.

The most effective processing path for determining a target component can be formed by ranking uncertainty values. By following this path and in any known condition, fewer candidates for the target component are left for choice compared to the other paths. If the target component needs to be recognized, the accuracy of recognition can be increased since the word recognizer or digit recognizer works better on a smaller size of lexicon [4]. It also means that a mail piece has higher possibility of being correctly interpreted in this case. Another advantage is that fewer components will need to be interpreted when the most effective path is used. The target component may be uniquely identified in the middle of the path. It indicates the advantage of improving processing speed and reducing error since fewer components need to be located and recognized. If a certain component cannot be located or recognized by applying the most effective path, an alternate component can be suggested by the measured information which is the second effective component to be recognized for a known condition. Hence, an alternate processing path is formed.

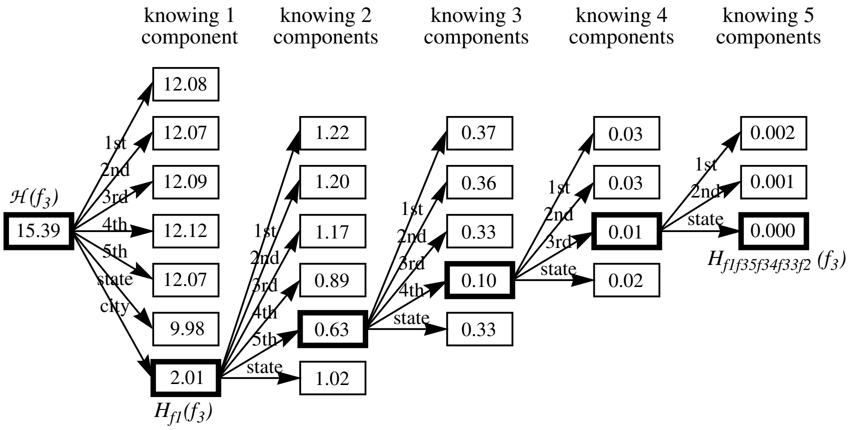


Figure 5. Most effective component to be recognized next for a known condition.

$\underline{\text{NY}}$	①	$\underline{4}$	$\underline{2}$	$\underline{2}$	$\underline{8}$
f_2	f_{31}	f_{32}	f_{33}	f_{34}	f_{35}

Figure 6. Illustration of recovering the 1st ZIP-Code digit from other components

Table 2. Redundancy measure

x	$\mathcal{H}(x)$	$\mathcal{H}(x, f_{31})$	$R_x(f_{31})$	Ranking
f_2	5.95	6.00	0.99	1
f_{32}	3.32	6.64	0	2
f_{33}	3.32	6.64	0	2
f_{34}	3.32	6.64	0	2
f_{35}	3.32	6.64	0	2
f_{31}	3.32	-	-	-

4.4. Recovering the first ZIP-Code digit from other components

If the confidence of a candidate for a component from a recognizer is low or close to that of other candidates, it is possible to recover this component by knowing other components (e.g., knowing state abbreviation to help determine the 1st ZIP-Code digit as shown in Fig. 6). The most valuable component for recovering a component can be discovered by measuring the redundancy of other components to this component. For f_2 and f_{3i} to f_{31}, the information measure result with the ranking order is shown in Table 2. Since f_2 possesses the largest redundancy quantity, it is the most valuable component to be recognized in order to verify the digit recognition result. From the database information, there are 62 state abbreviations. For 60 of them, the first ZIP-Code digit is unique. For the other 2 (i.e., NY and TX), there are 2 valid 1st ZIP-Code digits for each state abbreviation. The same scheme of redundancy measure can be applied to other fields of interest.

5. Conclusion

Information-theoretic analysis of postal address information has been shown to be helpful to automatic address interpretation. The efficiency of assignment strategies for determining a delivery point code can be compared by the information measure. A damaged component has higher possibility of recovery when another component with larger redundancy is known. From the measure of uncertainty of a target field, the most effective processing flow can be modeled. This scheme provides the processing flexibility for address interpretation.

Although examples provided are for US postal addresses, the analysis can be used for addresses of other countries. These presented schemes do not restrict the source of information. The information source can be from real mail stream or address directory. However, the information reliability and complexity of collecting information should be considered. Furthermore, if the information of processing complexity of a field in an address interpretation engine can be combined with these schemes, the effectiveness of the information can be enhanced.

References

[1] United States Postal Service, “1997 Annual Report,” *http://www.usps.gov/history/anrpt97/*.

[2] S. N. Srihari and E. J. Kuebert, “Integration of Hand-Written Address Interpretation Technology into the United States Postal Service Remote Computer Reader System,” the *4th ICDAR*, Ulm, Germany, 892–896, Aug. 1997.

[3] S. N. Srihari, “Recognition of Handwritten and Machine-Printed Text for Postal Address Interpretation,” *Pattern Recognition Letters*, 14(4):291–302, 1993.

[4] G. Kim and V. Govindaraju, “A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications,” *IEEE Trans. on PAMI*, 19(4):366–379, Apr. 1997.

[5] C. E. Shannon, “A Mathematical Theory of Communication,” *Bell System Technical Journal*, 27:379–423, July, Oct., 1948.