Dictionary Production for Census Form Conference

R. Allen Wilkinson

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Computer Systems Laboratory
Advanced Systems Division
Gaithersburg, MD 20899

NIST

Dictionary Production for Census Form Conference

R. Allen Wilkinson

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Computer Systems Laboratory
Advanced Systems Division
Gaithersburg, MD 20899

April 1993

Dictionary Production for Census Form Conference

R. Allen Wilkinson

National Institute of Standards and Technology,

Gaithersburg, MD 20899

Abstract

There are two categories of data from which dictionaries can be produced. One uses old data or data from a previous collection and the other uses new data or data from a current collection. The old data creates dictionaries that can be used for possible answer examples, assisting optical character recognition (OCR) systems, and training of recognition systems. The new data is the most useful in testing and scoring system results.

For each of the categories above there are two types of dictionaries. These types may be useful for work with the Second Census OCR Conference. The first contains all words that have occurred in the data set being used. For this experimental work, the data set is from the 1980 Census. These words can be misspellings, abbreviations, or correctly spelled words. This first or essential dictionary is easier to create and will not increase the errors which exist in the original data. This dictionary contains all the misspellings, abbreviations and other errors that occur when the original data was keyed from the original paper questionaires. This will make the dictionary useful in describing potential contents of a form set. The second dictionary can be built from the essential dictionary. The second dictionary is one which has the misspellings corrected, the abbreviations expanded, and all the words stemmed into logical minimal stems. A mapping between the essential dictionary to the second or exploratory dictionary is required. The exploratory dictionary is harder to create and may produce more errors than it corrects. This dictionary needs human assistance to be created since most of the steps can not be fully automated. It is also useful in showing a comprehensive list of the possible entries in a form set from previous data, in our case the 1980 Census.

Short and long dictionaries can be produced from both of these dictionaries. The long dictionary contains all of the data from the original Census data, while the short dictionary contains only the phrases and possibly words which occur more than once. The short phrase dictionaries are approximately 16% of the length of the long phrase dictionaries. They also contain 60% to 70% of the phrases in the original sample of 132247 phrases. The short word dictionaries are approximately 45% the size of the long word dictionaries. About 95% of the words found in the long phrase dictionaries are contained in these short word dictionaries.

1   Introduction

In May 1992, NIST hosted the First Census Optical Character Recognition (OCR) Conference. This conference tested the state of the art in OCR technology for isolated hand printed characters. The Second Census OCR Conference is going to address the issue of OCR on form images. The form images being used are "mini-forms". These "mini-forms" will be the Industry and Occupation sections of the 1990 Long Census forms, specifically questions 28b, 28c, 29a, and 29b of Form D-2. Each participant will be asked to recognize the hand printed entries for questions 28b, 29a, and 29b and return these results for scoring. The ability to use context to assist in the recognition process makes it desirable to produce dictionaries that contain the words or phrases likely to appear as entries.

There are two categories of dictionaries. The category depends on the type of data used to create the dictionary. The first category uses the data from the previous Census, 1980, and the resulting dictionaries will be used to give examples of possible answers and assisting OCR systems. The second category uses the data from the current Census, 1990. These dictionaries will be used for creating testing data and scoring data.

Within the above categories there are two types of dictionaries that can be produced and might be useful for this conference. The first contains word or phrase answers from similar questions in the Census data. These words include misspellings and abbreviations, as well as correctly spelled words. This type of dictionary is essential for scoring at the character level. This dictionary may also prove useful for improving word or phrase level OCR outputs. The second type of dictionary is one which has the misspellings corrected, the abbreviations expanded, and all the words stemmed into logical minimal stems. This type of dictionary can be built from the first type of dictionary. The second type may also be useful for scoring at the word level. A mapping between the essential dictionary and the second or exploratory dictionary is required. This allows for the creation of a new phrase list from the words in the exploratory dictionary. This new list will still represent the data from the original list but will have consistent phrasings.

2   Producing the Essential or First Dictionary

The data starts as a key punch operator's view of what was entered in the Industry and Occupation fields of the 1980 or 1990 Long Census Forms. A sample of 132247 keyed responses is used with each form having four lines of Industry and Occupation data. The first line, Line 0, is a code for the Industry and Occupation information which follows. The other three lines, Line 1, Line 2 and Line 3, are phrases. These phrases need to be processed to produce a dictionary. The process is to change form entries to phrases, then the phrases to words. The final result is a dictionary that contains all the words which occur in the phrases. This dictionary is long and contains many spelling errors and abbreviations.

2.1 Turning form entries into 3 different phrases

Each form entry has one encoding line and three text lines. The first line contains Census encoding for the data that follows. For the dictionary production it is not important. Line 1 is the response to, "What kind of business or industry was this?", question 28b. Line 2 is the response to, "What kind of work was this person doing?", question 29a. Line 3 is the response to, "What were this person's most important activies or duties?", question 29b. It is easy to see how the three lines can be interrelated but each line is viewed separately in the dictionary creation process. The very small list in Figure 1 will be used to demonstrate the process of dictionary creation. The phrase lists are generated by removing the same line from all forms. For instance, one phrase list will contain all the Line 1 responses. All punctuation is removed from these phrases. Figure 2 lists all the Line 1 phrases for the short sample list with punctuation removed. The list of phrases from the original Census data has many entries that have multiple consecutive spaces. These are converted into one space to clean up the phrase list. This list is then sorted in alphabetical order and all duplicate entries are removed. A small set of real Line 1 phrases from the original 1980 Census entry list are shown in Figure 3. The resulting lists are very large. For Line 1 there are 46593 unique entries, while Line 2 has 46813 and Line 3 has 61384.

```
 0  0  4  1  412 354
ZNT-OPERATOR
OPEATOR
OPERATOR
 0  0  1  1  250 259
GLASS MARUF
MANUFACTURING GLASS
MANUFACTURING GLASS
 0  0  4  1   11 274
FFED LOOT
SELLING CATTLE
SUPERVISOR
 0  0  0  1  391 674
LAMPSHADE MARV
MAKING LAMPSHADE
GIVING
 0  0  4  2  910 179
BKING JUDGE
JUDICAL
JUDGE
```

Figure 1: The entries of the demonstration sample list

```
BKING JUDGE
FFED LOOT
GLASS MARUF
LAMPSHADE MARV
ZNT OPERATOR
```

Figure 2: Examples of phrases found as entries from Line 1 of the demonstration sample list after removal of punctuation and sorting

2.2   Making Phrases into Words

The conversion of phrase lists into word lists is rather simple. The first step is to remove all entries that might violate the Privacy Act, such as addresses. Then, using UNIX utilities, it is possible to replace all spaces with newlines. This puts each word on a line of its own. The lists are once again sorted in alphabetical order and duplicate entries are removed. The word list for Line 1 has 13745 words, Line 2 has 13879 words, and Line 3 has 16333 words for this experiment. These numbers show an extreme decrease in list size. This can be attributed to the redundancy of words within the phrase lists.

Figure 4 shows the words produced from the phrases in Figure 2. Figure 5 shows a small sampling of the words from the real Line 1 phrase list. Notice both of these examples contain misspelled words. These lists may also contain words which have the same stem. For example, "advisement" and "advising" have the same stem as "advisor", "advisors", and "advisory". That stem is "advise".

2.3   Problems with the First Dictionary

The First dictionary is appealing because it significantly reduces the dictionary size compared to the size of the full 1980 Census data list. Unfortunately, there are also problems with this dictionary. The problems include misspellings, abbreviations, and erroneous data entry. The erroneous data entry was noticed because the key punch operator was to enter the word "blank" when a blank field was found. The word "blank" was misspelled in many ways. Another data entry problem is key punch operators may sometimes spell correctly what is misspelled on the original form. These errors will be discussed in section 3.

3   Production of Second Dictionary from First

The second dictionary needs to be produced in order to improve the scoring and reduce errors. The type of scoring used will help determine if the second dictionary is necessary.

```
ACADEMIC INSTUTUTION
ACADEMIC LIBRARY BOOK JOBBER
ACADEMIC PEDIATRIC PRACTICE
ACADEMIC PHYSICS
ACADEMIC RESEARCH
ACADEMIC RESEARCH CENTER
ACADEMIC SCIENCE DEPTS
ACADEMIC TEACHING
ACADEMIC UNIVERSITY
ACADEMIC ZOOLOGY RESEARCH TEACHING
ACCESS CONTROL MFG
ACCESS FLOOR SERVICE CENTER
ACCESSIBILITY SURVEYOR ANALYST
ACCESSORIES FOR KNITTING HILLS
ACCIDENT INSURANCE FIRM
ACCOOPIONAL TABLES
ACCOUNBEE HEATING
ACCOUNT
ACCOUNT PAYABLE DIVISION
ACCOUNT REP
```

Figure 3: Examples of real phrases found as entries in the 1980 Census sample

Do you score on *what exists on the image* or by *what the writer intended to convey*? The best choice is to score the results on *what exists on the image* because that is what was requested of the data entry personnel. To achieve this, the image must be processed accurately by a human. It is known that this did not happen in all cases with the Census data and therefore the second dictionary is necessary.

One problem is key punch operators misspell words, as mentioned in section 2.3. The word "blank" was to be entered for every field that was blank. Blank is misspelled several possible ways in this data: balnk, blnak, blak, blan, blanks, and blasnk. Therefore, misspellings created by a key punch operator exist. Table 1 shows the occurrence rates of the possible spellings for the word blank. These problems can be solved by creating a correctly spelled dictionary. It must be understood that as soon as you create a correctly spelled dictionary, you can not score by *what exists on the image*.

Another not so obvious problem is key punch operators can read semantically. Most humans, when entering information into a system, read what was written at the word level. When doing this, some misspelled words are close enough to the proper or correct spelling that humans can recognize the word. The word is then spelled correctly when entered. We have no data about the rate of occurrence of this type of error. This kind of error suggests the use of *what the writer intended to convey* scoring. To allow for the fact that we do not always know what word was intended with misspellings, the correctly spelled words are also stemmed. The intended word interpretation can be an even greater problem with abbreviations. One added benefit from stemming is the reduction

```
BKING
FFED
GLASS
JUDGE
LAMPSHADE
LOOT
MARUF
MARV
OPERATOR
ZNT
```

Figure 4: Examples of words extracted from phrases of the demonstration sample list

of the dictionary size. This allows for many entries: misspelled, abbreviations, and extensions of the stem, to be mapped to one word. However, this process introduces problems as discussed below.

3.1 How is the first type of dictionary processed into the second type

This section explores the issues associated with the production of dictionaries of intended responses. This allows scoring of the word level as an alternative to scoring at the character level where errors in the keyed responses are known to exist. One motivation for producing such a dictionary would be to provide more realistic scores at the word and field level than by simply requiring all characters in a word or field to be correct. It does not appear correct to give the same word and field level scores to "brake manufacturer" and "abrasive manure tower" when the keyed reponse was "brakes manufracturer".

The initial processing of the first type of dictionary is to convert it to lower case. An upper case word will cause strange side effects in the spelling and stemming utilities used. Then the words are spelling corrected and stemmed. These stems are then inserted back into the original phrases to produce correctly spelled and stemmed phrases. These last three steps of the production; correcting spelling, stemming, and making stemmed phrases, need human assistance. This may introduce word level errors while reducing character level errors introduced by the key punch operators.

3.2 Making words into correctly spelled words

The word list must be passed through a spell checking program. An operator must look at the flagged words and decide whether to keep the spelling correction or not. Some of the words make no sense. Some words are abbreviations and need to be expanded.

```
ADVICORY
ADVISEMENT
ADVISING
ADVISMENT
ADVISOR
ADVISORS
ADVISORY
ADVOCACY
ADVOCATE
ADVRTISING
AENCY
AEOROSPACE
AERATOR
AERESOL
AERIAL
AERO
AEROBIC
AEROCPACE
AERONAUTICAL
AEROPLANE
```

Figure 5: Examples of words extracted from phrases in the real list of phrases for the 1980 Census sample

Figure 6 shows the misspelled words from the demonstration sample list. Figure 7 shows some of the real misspelled words or abbreviations which need to be corrected. Some words from this list can be corrected while some of them require more information. It is possible to look back at the phrase list to resolve ambiguities. This is part of the reason that mapping between dictionaries, as discussed earlier, must be maintained. Examples of these questionable words are shown in Figure 8 with misspelled words in lower case followed by the phrase in upper case. This figure shows that it may be helpful to look back at the original form entry file, but that even then it may not be possible to resolve all ambiguities. At this point, we have not tried to correct spelling errors, or to expand abbreviations.

```
bking
ffed
lampshade
maruf
marv
znt
```

Figure 6: Examples of misspelled entry words of the demonstration sample list

Word as found in Entries	Occur in Phrase	Percentage Occurred	Occur as Word	Percentage Occurred
BLANK	30457	97.469%	30499	92.539%
BANK	714	02.285%	2091	06.344%
BALNK	20	00.064%	21	00.064%
B LANK	8	00.026%	8	00.024%
BLNAK	6	00.019%	6	00.018%
BLNK	4	00.013%	4	00.012%
CLANK	3	00.010%	3	00.009%
BAND	2	00.006%	145	00.440%
BLAN	2	00.006%	12	00.036%
BLASNK	2	00.006%	2	00.006%
BLAK	2	00.006%	2	00.006%
BLOCK	1	00.003%	136	00.413%
BLANKK	1	00.003%	1	00.003%
BALANK	1	00.003%	1	00.003%
BLANKB	1	00.003%	1	00.003%
BLSANK	1	00.003%	1	00.003%
BLAMK	1	00.003%	1	00.003%
BLSNK	1	00.003%	1	00.003%
BNAK	1	00.003%	1	00.003%
BLANKS	0	00.000%	1	00.003%
total	31248	100.000%	32958	100.000%
total misspelled	791	02.531%	2459	07.461%

Table 1: Table of Occurances for Blank and its misspellings.

3.3   Making stems

The UNIX operating system has a spell checker, *spell -x*, that does stemming. Stemming is the removal of suffixes and/or prefixes from a word to produce a shorter word, a stem. The root of a word is the shortest stem possible for that word. A word can have many stems but only one root. In Figure 10, the root for the word "absorbers" is absorb while there are several stems. The appropriate stem must be found and it is not always the root. An operator must decide which stem is appropriate.

Examples of real stemming can be seen in Figure 10, while the stemmings for the demonstration sample list are in Figure 9. When stemming is applied, the word "biomedical" becomes "medic" for a root but the stem "medical" may be more appropriate. Also, "registers" becomes "gist", which is probably not acceptable.

3.4   Making stemmed phrases

Stemmed phrases can be made by mapping stems to their correctly spelled words. Then these words are mapped to the word lists. The stem can then be substituted for

```
afair
aric
bking
bunrering
cartiage
cig
divistion
eledry
eretioin
ffed
horing
knouse
lanoman
litho
maruf
marv
mazine
peoria
plub
steaqm
tren
varn
zipvelope
znt
```

Figure 7: Examples of misspelled entry words

the word in the original phrases. This step also needs human inspection to make sure the new phrases are correct. This step has not been attempted yet.

4 Other Facts

All the dictionaries described so far are complete and long dictionaries. This means every word from the original entries is processed and represented in the final result. Short dictionaries can be made which will decrease the length of the word lists. These dictionaries are made by keeping only phrases which occur more than once in the entry lists. The length of the phrase lists then becomes 8216 entries for Line 1, 8516 for Line 2, and 7831 for Line 3. This is an extreme reduction in the list sizes, while still giving good coverage of the original list of phrases. The lists are 15% to 20% of the long list size, but represent 60% to 70% of the phrases in the original forms. It also appears that the addresses found in the long lists are not present in the short lists.

The short word lists are generated by using the long phrase list and keeping the words that occur more than once. The sizes are 5950 words for Line 1, 6014 words for Line 2,

```
afair CITY AFAIR
aric CHEMICALS ARIC
bking BKING JUDGE
bunrering OIL SALES CRUISE BUNRERING
cartiage CARTIAGE
cig CIG MANUFACTURE
divistion DIVISTION OF CORRECTIONS
eledry ELEDRY HOME
eretioin STEEL ERETIOIN
ffed FFED LOOT
horing MACHINE SHOP DRI HORING SOLAR PA
knouse GUARD DUTY FOR KNOUSE FOODS
lanoman PETROLEUM LANOMAN
litho LITHO TRADE SHOP
maruf GLASS MARUF
marv LAMPSHADE MARV
mazine MAZINE BRUDRY BOOKS
peoria DOOR WORK PEORIA ILL NUCLEAR PLAN
plub AUTO SPARK PLUB MFG
steaqm STEAQM CARPET CLEANING
tren PUETTAS PARA TREN
varn MANUFACTURING VARN
zipvelope ZIPVELOPE MFG CO
znt ZNT OPERATOR
```

Figure 8: Examples of misspelled entry words and their associated phrases

and 7605 words for Line 3. This is approximately 45% of the size of the unique word lists from the long phrase dictionaries. These short word lists contain about 95% of the words found in the long phrase dictionary. Further investigation must take place before it can be determined if this size reduction will have any adverse effects on the dictionary's usefulness.

```
glass   =glass
judge   =judge
loot    =loot
operator  =operate =operator
```

Figure 9: Examples of stemming on the demonstration sample list

5 Conclusions

There are two categories of data from which dictionaries can be produced. One uses old data or data from a previous collection and the other uses new data or data from a current collection. The old data creates dictionaries that can be used for possible answer examples, assisting OCR systems, and training of recognition systems. The new data is most useful in testing and scoring system results.

For each of the catagories above there are two types of dictionaries which can be produced from Census data. The first, the essential dictionary, is easier to create and will not produce more errors than exist in the original data. This dictionary contains all the misspellings, abbreviations and other errors that occurred when the original data was keyed from the images. The essential dictionary is useful in describing the potential content of a form set. It is also useful for scoring at the character level.

The second, the exploratory dictionary, is harder to create and may produce more errors than it corrects. The second dictionary is one which has misspellings corrected, the abbreviations expanded, and all the words stemmed into logical minimal stems. This dictionary needs human assistance to create, since most of the steps can not be fully automated. The exploratory dictionaries are useful for scoring at the word and field level. They would contain only corrected entries that make them much smaller than the essential dictionaries.

Short and long dictionaries can be produced from both of these dictionaries. The long dictionary contains all of the data from the original Census sample data, while the short dictionary contains only the phrases and possibly words that occur more than once. The short phrase dictionaries are approximately 16% of the length of the long phrase dictionaries, but contain between 60% and 70% of the phrases in the original sample of 132247 phrases. The short word dictionaries are approximately 45% the size of the long word dictionaries. At the same time the short word dictionaries contain about 95% of the words found in the long phrase dictionary.

```
absorbers =absorb =absorbe =absorber =absorbers
accessoriers =accessory =accessorie =accessorier =accessoriers
activities =act =action =active =activity =activitie =activities
biomedical =medic =biomedic =medical =biomedical
computerizing =compute =computer =computere =computerize =computerizing
discount =count =discount
electroplating =plate =electroplate =plating =electroplating
engeneering =gene =engene =genee =engenee =geneer =engeneer =geneere
            =engeneere =geneering =engeneering
engineering =engineer =gineere =engineere =gineering =engineering
geologists =geology =logist =geologist =logists =geologists
havester =have =havest =haveste =havester
lettering =let =lett =lette =letter =lettere =lettering
microbiological =logic =biologic =microbiologic =logical =biological
                =microbiological
mispant =pant =mispant
missiles =missile =siles =missiles
monogramming =gram =gramm =monogramm =gramming =monogramming
organizational =organ =organe =organize =organization =organizational
outfitters =fit =fitt =outfitt =fitte =outfitte =fitter =outfitter
            =fitters =outfitters
overhauling =haul =overhaul =haule =overhaule =hauling =overhauling
preparedness =prepare =pared =prepared =paredness =preparedness
rebuilders =build =rebuild =builde =rebuilde =builder =rebuilder =builders
            =rebuilders
receptionist =reception =ceptione =receptione =ceptionist =receptionist
registers =gist =regist =giste =registe =gister =register =gisters =registers
semiconductors =conductor =semiconductor =conductors =semiconductors
subcontracting =contract =subcontract =contracting =subcontracting
subsidized =side =subside =sidize =subsidize =sidized =subsidized
supervision =vision =supervision
supplemental =supple =supplement =supplemental
undergarments =gar =undergar =garment =undergarment =garments =undergarments
underwriters =write =underwrite =writer =underwriter =writers =underwriters
unemployment =employ =unemploy =employment =unemployment
vacations =vacate =vacation =vacations
winchester =winch =winche =winchest =wincheste =winchester
zipper =zip =zipp =zippe =zipper
```

Figure 10: Examples of stemming