AN EXPERIMENTAL SYSTEM FOR AUTOMATIC RECOGNITION OF PERSONAL TITLES AND PERSONAL NAMES IN NEWSPAPER TEXTS

Casimir Borkowski

Thomas J. Watson IBM Research Center

Yorktown Heights, N.Y.

Summary.

Natural language seems to contain various special-purpose sublanguages (e.g., personal titles, personal names) -- each with its own structure which relative to the total structure of language is quite simple.

An ability to generate and to recognize automatically words and word strings belonging to various special-purpose sublanguages may prove to be very useful since they play an important role in indexing and in various systems for extracting and distributing information.

This paper (1) describes some of the main problems involved in automatic recognition of personal titles and names in newspaper texts, (2) outlines some rules of an algorithm designed to perform this task, (3) presents statistics concerning the algorithm's accuracy and exhaustiveness obtained in manual application of the algorithm to texts, (4) discusses and interprets some of the results, and (5) suggests some applications for computer programs capable of recognizing personal titles and names.

***

Motivation for the Experiment.

One of the major questions of the day is the extent to which a computer can be instructed to identify various parts of texts written in plain, ordinary language. In trying to answer this question, we set ourselves the preliminary limited objective of developing an automatic procedure for identifying personal titles and personal names in English-language texts.

Experimental Design.

Our procedure in setting up an automatic method for identifying personal names and titles was approximately as follows:

(1) We investigated permissible patterns of personal titles and of English, French, Russian, German, Spanish, Chinese, Arabic, and other personal names whose occurrence in texts we could anticipate.

(2) We obtained a 60,000-word sample of newspaper texts and determined: (a) patterns of occurrence of personal names and titles in texts, (b) patterns of personal names and titles occurring in texts, and (c) problems involved in distinguishing personal names and titles from each other and from other parts of texts.

(3) Based on (1) and (2) above, we set up an automatic procedure designed to identify personal names and titles in newspaper texts. This procedure was embodied in flowcharts and a dictionary of about 8,000 entries..

(4) We tested our procedure manually on a 100,000-word sample of new newspaper texts, and we amended the rules and expanded the dictionary on the basis of the information provided by the tests.

(5) We then stabilized the improved procedure and (a) tested it out manually on a new 40,000-word sample of newspaper texts and (b) collected statistics concerning its accuracy and exhaustiveness. (Our reasons for applying the algorithm manually were as follows: (a) our identification system was embodied in dictionary entries and flow charts which were sufficiently detailed to permit accurate execution of recognition procedures, and (b) we thought that it would not pay to code and debug over a period of months what would probably turn out to be a "one-shot" program.)

(6) We then investigated what types of errors had occurred and proposed various amendments to the automatic recognition procedure.

Some Problems of Automatic Identification of Personal Titles and Names.

Automatic identification of titles and names in texts is of course not without its difficulties. First of all, many personal names are orthographically identical with other types of words in the language. This is the case since among the main sources of surnames are: (1) titles (e.g., "King"), (2) names of occupations (e.g., "Baker"), (3) topographic terms (e.g., "Hill"), (4) personal attributes (e.g., "Coward"), (5) place names (e.g., "London"), (6) names of animals (e.g., "Fox"), (7) names of trees (e.g., "Pine"), etc.

There is considerable ambiguity between personal names and place names due to the fact that not only are the names of places a frequent source of personal names but also because many localities were named after people, as for example, Elizabeth, New Jersey andDallas, Texas. And to make matters worse, hotels, business firms, universities, etc. can be named after people and are often referred to by an abbreviated name which is that of a person (e.g., "He is staying at the Hilton", "He graduated from Stanford", "Ford was hit by a strike last week"; "Two of them climbed the Everest"). As for personal names like "Helena Rubinstein" and "Max Factor", they designate persons as well as business firms, while "Philip Morris" is the name of a person, of a corporation, and of a brand of cigarettes.

Yet another difficulty arises in case of names of persons (e.g., "Madison") when they perform a naming function with regard to something, say an avenue, (e.g., "Madison Avenue"). Presumbly, it would be worthwhile to distinguish automatically references to persons from references to things named after persons.

Further difficulties in automatic recognition result from the co-occurrence in texts of names belonging to different name strings (e.g., "John Byron" as in "Estelle gave John Byron's Don Juan", "Mary Jane" as in "For Mary Jane had nothing but sympathy", "Alexander Montgomery" as in "According to Alexander Montgomery was slow in exploiting successes").

Other difficulties in recognizing personal names result from the fact that personal titles are not unfailing aids in identifying and disambiguating personal names since titles themselves can be homographic with other types of words. For instance, "General" is a military rank in "General Mobutu", but not in "General Motors".

Further difficulties result from the fact that some titles are homographic with given names. How is an automaton to tell that "Dean" is a title in "Dean Wiesner" but a name in "Dean Rusk", that "King" is a title in "King James" but a name in "James King", that "Earl" is a title in "the fourth Earl Russell" but a name in "the Chief Justice Earl Warren"?

Some Recognition Rules.

Our recognition algorithm was intended as a frame of reference in an investigation of the trade-off between the efficiency and the complexity of a series of algorithms.

To be able to investigate the trade-off between the efficiency and the complexity of a series of algorithms, we avoided taking as our point of departure a strong theory about the structure of language, semantics, pragmatics, etc., and about the amount of syntactic recognition required for successful identification in texts of personal titles and names. Instead, we sought to discover what assumptions and what information about natural language and texts may be pertinent to the resolution of the limited problems in recognition which we set for ourselves.

Stronger assumptions and more elaborate techniques of analysis can be built into subsequent algorithms if required and as required. For instance, since parsing may be helpful in identifying sequences of names each of which is followed by its title (e.g., "The President nominated John Gordon Ambassador to Guatemala, William T. M. Beale Jr. Ambassador to Jamaica...") future recognition algorithms may parse sentences containing: (1) double-object verbs (e.g.,"nominate") and (2) strings consisting of personal names followed by titles.

At a later time, parsing and/or other types of analyses may be extended to sentences, paragraphs, and articles containing other kinds of words and phrases. However, since parsing and other types of analyses may be expensive, it would seem advisable to apply them only when they can reasonably be expected to provide economic solutions to valid problems.

Our rules describe the arrangement in the sentence of the words, phrases, and punctuation marks which are pertinent to recognition of names and titles. Generally, the description starts with the first, that is, the leftmost pertinent element and terminates with the last, or rightmost pertinent element. Recognition rules were given a "left-to-right" format because rules expressed in this way are easy to implement on an electronic computer.

For greater ease of understanding, the rules are expressed here in narrative form. For the sake of brevity, only some recognition rules are listed here. A more complete description of identification rules is available elsewhere.(1)

Our rules for recognizing names of persons take advantage of the style rules of The New York Times. We would conjecture that whereas details of name recognition rules may vary from newspaper to newspaper, their general pattern will remain fairly stable and independent of editorial conventions.

The rule for identifying personal titles which was selected as a reasonable first approximation states that a word or phrase in text is a personal title either:

(1) if it matches a word or string of words on a list of titles

or

(2) if it matches a word or a string of words which is on a list of words and phrases which commonly combine with titles (e.g., "Acting", "Assistant", "Vice") and is followed by a personal title (e.g., "Acting Mayor", "Acting Assistant Vice President")

or

(3) if it is a personal title followed by a word or a string of words which is on a list of words which commonly combine with titles (e.g., "-elect", " at Large", "pro tempore") as in "Senator-elect", "Ambassador at Large", "President pro tempore".

or

(4) if it is a title designated by a list, like say "Commissioner", followed by the word "of" and any capitalized word (e.g., "Commissioner of Parks").

or

(5) if it is a word beginning with a capital letter and followed by a title designated by a list (e.g., "Commissioner" as in "Police Commissioner").

A preliminary (and a highly tentative) rule specifies how titles concatanate. This rule permits distinguishing some strings such as "Prime Minister, Sir" as in "Prime Minister, Sir Alec Douglas-Home", "Rev. Dr." as in "Rev. Dr. Martin Luther King", "Mr. Chairman"," Mr. Counsel", "Mr. Chairman, Ladies, and Gentlemen:", and so forth from titles followed by names.

The present set of rules for identifying titles which are homographic (that is, ortographically identical) with other words is relatively simple. It divides ambiguous titles into four classes; words of Class I (e.g., "King", "Pope", "Prince") are assumed to be titles (e.g., "King John", "Pope John") unless:
(1) preceded by either personal titles designated by a list such as "Mr.", "Dr.", "M. Sgt.", "General", etc., or by given names and initials in various combinations (e.g., "Mr. King", "John King", "Dr. Pope", "John Pope", "John M. King", "J. M. King")

or

(2) followed by such postnomial elements as "Sr." (e.g., "King, Sr."), "& Bros." (e.g., "King & Bros.), "and Company" (e.g., "King and Company"), and so forth.

Occasionally, words of Class I are followed by capitalized words or phrases which designate various institutions, establishments, locations, and so forth which are frequently named after persons (e.g., "Drug Store", "College", "Avenue", "Theorem"). Although, the list of such words and phrases is open-ended, its most frequently occurring members can be discovered and listed quite easily. Furthermore, there is some evidence that many or most such phrases can be identified by means of recognition rules. Words which are members of Class I are assumed to be names when they are followed by words such as "College" or phrases such as "Drug Store".

Words which are members of Class II (e.g., "Kaiser", "Chamberlain", "Earl") are assumed to be personal names.. Commonly occurring exceptions to this rule (e.g., "Kaiser Wilhelm", "Lord Chamberlain", "Earl of" (if followed by a word beginning with a capital letter)) are listed.

Words which are members of Classes III and IV (e.g., "General", "Principal", "Justice") are assumed to be titles. Commonly occurring exceptions to this rule are listed ("General Assembly", "Major Medical Plan", "Principal Investigator", "Justice Department").

As our rules become more sophisticated, the need for lists of exceptions will diminish. However, it is likely that listing exceptions will often be an attractive alternative to rendering a rule more complicated.

Personal titles in the plural are recognized by means of a simple rule. It states that a string of characters is a personal title in the plural if:

(1) it is recognizable as a personal title

and if either

(2) its final word is followed by the letter "s" (e.g., "Major Generals")

or else

(3) if one of the words of which it is composed and which a list designates as the stem for the plural is followed by the letter "s" (e.g., "Collector" as in "District Collector of Internal Revenue").

A procedure similar to the one for identifying titles in the plural is used to recognize personal titles in the possessive case (e.g., singular: "Major General's", plural: "Major Generals'").

Our rules assume that the capitalized word or string of words and initials which frequently follows a title is the name of a person (e.g., "President Nkrumah", "Mr. Paul-Henri Spaak", "Governor Nelson Rockefeller").

If a title is followed by a word beginning with a lower-case letter or by certain punctuation marks, this indicates that the title is not followed by a name. However, occasionally personal titles are followed by names beginning with lower-case letters (e.g., "President de Gaulle").

On occasion, titles are followed by capitalized words which are not names. This happens in particular when a title is followed by a capitalized word or phrase which designates an institution, an establishment, a site, and so forth, named after a personal title (e.g., "Ambassador Bar", "Archduke Trio", "Emperor Concerto", "President Hotel", "Queens County", "Viceroy Lumber Company").

Although -- as mentioned earlier -- the list of such words and phrases is open-ended, its most frequently occurring members have been listed and are consequently identifiable. Commonly occurring exceptions to this rule are also listed and are therefore identifiable. In addition, we have some simple preliminary rules for identifying phrases whose designata often bear as names words or phrases which are personal titles (e.g., "President Radio Repair Shop", "Viceroy Lumber Company"). However, since the identification of phrases which designate such namesakes has been given little attention, these rules are very tentative.

Words which are generally names of weekdays when they occur after personal titles (e.g., "They saw the President Monday") are of course listable and therefore identifiable.

Occasionally, titles are followed by capitalized words which are not names and for the recognition of which the rules make no provisions (e.g., "British" in "The Prime Minister, British sources said, will arrive on Monday." and "New York" in "Mr. Stevenson prefers Washington and Mr. Rusk, the Secretary of State, New York."). Constructions such as these are, however, quite rare.

Occasionally, prepositional and other phrases intrude between a title and a name (e.g., "the French Ambassador to the United States, Herve Alphand", "the Foreign Minister of France, Maurice Couve de Murville"). Prepositional phrases of this sort and other adjuncts are identified by various rules of the "brute force" type whose statements are constructed as follows: If a personal title (e.g., "Foreign Minister") is followed by a phrase consisting of the preposition "of" and of the name of a country, such phrase is part of the title.

In general, in a string of words consisting of titles and names, titles precede names (e.g., "the Secretary of State, Dean Rusk, the Foreign Minister of the Federal Republic, Gerhard Schröder, the Foreign Minister of France, Maurice Couve de Murville"). Sequences of names each of which is followed by its title (e.g., "Dean Rusk, the Secretary of State, Gerhard Schröder, the Foreign Minister of the Federal Republic, ...") are rare. (Ordinarily, in a construction of this type, each title is set off from the name which follows it by a semi-colon.)

In spite of counter-examples such as the ones above, one can reasonably assume that if the capitalized word or string of words and initials which frequently follows a personal title is NOT an identifiable word like "Hotel", "Garage", "Street", "Monday", etc., or a phrase like "Barber Shop", "Drug Store", "Meat Packing Company", etc., then it is a personal name.

While counter-examples to this rule and to similar rules are easy to invent, the inventors of counter-examples usually miss the point that rules such as these are statistical observations and that in actual application to texts they hold up rather well. As stated earlier, among the goals of an investigation of this type is to obtain experimental evidence as to how well the rules hold up and what amendments are required to simplify them, to render them more accurate, and to expand their scope.

If a personal title (e.g., "President") is conjoined to the titles "Mrs." or "Miss" (as in "the President and Mrs. Johnson"), the capitalized word or string of words which follows the conjoined titles is generally the name of a person.

In general, the name of a person acts distributively with regard to the preceding titles, that is to say, a phrase like "the President and Mrs. Johnson" decomposes into "President Johnson and Mrs. Johnson".

Occasionally, a personal name does not act distributively with regard to conjoined titles which precede it (as in "an agreement between the Cardinal and Mrs. Johnson", "a meeting between the President and Mrs. Luce"); however, the present set of rules makes no provisions for recognizing such cases.

Generally, in newspaper articles, capitalized words and initials which (a) frequently follow a personal title in the plural (e.g., "Senators") and (b) which are not followed by other titles (e.g., "Senators, Congressmen, and Generals") are strings of personal names (e.g., "Senators Javits and Kennedy", "Senators Jacob Javits, Robert Kennedy and George D. Aiken", "Presidents Johnson and Lopez Mateos").

Of course, "Ambassadors Bar and Grill" is not a title in the plural followed by two names. However, commonly occurring phrases such as "Bar and Grill" are listable and therefore identifiable.

As a rule, the first name string -- which is often separated from the title by a comma -- terminates before the first conjunction "and" or before the next comma; the second name string begins after "and" or after the comma; etc.

If two name strings which follow a title in the plural (e.g., "Senators Jacob Javits, Robert Kennedy, ...") are separated by a comma, then -- generally -- the end of the second name string is marked by a comma or by the conjunction "and".

If two name strings which follow a title in the plural are separated by "and", then -- generally -- the end of the second name string is marked either by punctuation marks such as sentence period, a colon, a semi-colon, etc., or by a word beginning with a lower-case letter (e.g., "arrived" in "Presidents Johnson and Lopez Mateos arrived today."). However, if the word beginning with a lower-case letter is a name conjunction (e.g., "de", "von"), then such word does not mark the end of the second name (e.g., "Presidents Lyndon Johnson and Charles de Gaulle").

Occasionally, the string of words which follows a title in the plural consists of both names and prepositional phrases (e.g., "Senators Javits of New York and Fulbright of Arkansas", "Senators from New York, Javits and Kennedy"). Our rules permit identifying some prepositional and other phrases which may intrude between titles in the plural and the names which follow them.

Generally, the title in the plural acts distributively with regard to the names which follow it, that is to say, a phrase like "Senators Javits and Kennedy" decomposes into "Senator Javits and Senator Kennedy".

Generally, the end of a name string which may follow a personal title in the singular is marked by punctuation (comma, sentence period, dash, semi-colon, colon, exclamation point, apostrophe, three dots, left or right parenthesis, etc.) or by a word beginning with a lower-case letter.

However, a lower-case letter does not mark the end of a name string if the word which begins with it is either:

(1) a name conjunction (e.g., "de" as in "Attorney General Nicolas deB. Katzenbach")

or

(2) the last element of a hyphenated Chinese given name (e.g., "lai" in "Premier Chou En-lai")

or

(3) the one-letter Spanish word "y" (e.g., "President Jose Bustamante y Rivero")

or

(4) if it is one of the Arabic words "ibn", "el", "al", "er", and so forth (as in "Abdul-Assiz ibn-Saud", "Abd-el Kader", "Abd-al-Kadir", "Abd-er-Rahman").

The end of a name string is often marked by its last element (e.g., "Jr." as in "Rev. Dr. Martin Luther King Jr.", "2nd" as in "Douglas MacArthur 2nd", the Roman numerals "I", "II", "III", etc., as in "King Idris I"). Cases in which Roman numerals are in the middle of a name are rare and can be treated as listable exceptions (e.g., "King Gustaf VI Adolf").

<u>NOTE:</u> The present rule makes no provisions for distinguishing Roman numerals "I", "V", and "X" from the first person pronoun "I" and from the letters "V" and "X" since contexts in which these ambiguities may cause error in name recognition seem rare (e.g., "Malcolm X", "Pope Leo X", "Idris I", "May I leave?").

Ordinarily, a left parenthesis or a left bracket are among the punctuations which mark the end of a name string. This, however, is not the case when a person's title and given name are followed by his nickname in quotation marks, as for example in "Gen. Howell ("Howling Mad") Smith", "Adm. William ("Bull") Halsey", etc.

Similarly, whereas ordinarily a left bracket marks the end of a name string, occasionally, when quoting someone, newspapers supply in brackets the part of name which the original statement omitted (e.g., "My agreement with Senator [Richard B.] Russell..."); sequences such as these do not mark the end of a name.

The rule for identifying nicknames which was selected as a reasonable first approximation states that "strings of words in parentheses and quotes which occur immediately after the title and/or given names and before a surname are nicknames".

A parallel rule serves to identify names in brackets which act as amplifications of original statements.

The preceding section has stated in considerable although by no means full detail some rules for identifying titles and names. We hope that this form of presentation indicates the vast amount of detail involved in rules for automatic recognition without, however, overburdening the reader with a multitude of minute points of information.

Results of the Experiment.

Since our identification rules were embodied in dictionary entries and flow charts which were sufficiently detailed to permit an accurate manual execution of identification procedures, it was decided that our identification system would be tested out by hand on a sample of The New York Times texts.

Identification procedures were applied manually to some 40,000 words of texts. Altogether eighty-eight articles from eleven issues were selected and processed. Only newsarticles were included in the sample. All materials found in the special sections such as (1) entertainment, (2) food-fashions-family-furnishings, (3) social events, (4) necrology, etc. were omitted. Materials in the sample consisted of only texts of newsarticles; picture captions, advertisements, italicized lists of various sorts, charts and diagrams, etc. were excluded from the data.

Our 40,577-word sample contained 806 occurrences of names of persons. Of the 806 occurrences of names of persons, 46 or about 6% of the total were missed. In addition, 47 words and word strings were mistakenly identified as personal names or personal titles.

Figure of merit F for the results of this identification system was computed by means of the following formula:

$$F = \frac{C^2}{(C + M) \text{ x } T}$$

where C is the number of correct identifications, M is the number of mistaken identifications, and T is the number of names of persons in the sample. (2)

For T = 806, C = 746, and M = 47

$$F = \frac{746^2}{(746 + 47) \text{ x } 806} = .87$$

Analysis of Major Errors.

Twenty-six misses (out of a total of forty-six) and thirty mistaken identifications (out of a total of forty-seven) occurred in attempted identifications of words, word stems, and word strings which perform a naming function vis-a-vis some namesake (e.g., "Grumman Aircraft Engineering Corporation"). This source of misses and false identifications would be eliminated if in the future the automatic identification system was not required to decide whether words, word stems, and word strings (e.g., "Grumman") performing a naming function vis-a-vis some identifiable namesake (e.g., "Aircraft Engineering Corporation") are names of persons.

We also need more effective rules for computing namesake phrases (e.g., "Aircraft Company") and personal titles (e.g., "Fireman Apprentice") from their respective elements (e.g., "Aircraft", "Company", "Fireman", "Apprentice").

In addition, we need to prevent or eliminate the errors caused by the assumption that all capitalized words occurring after ambiguous words such as "General", "Justice", "Major", "Principal", etc. are names of persons.

We also require more effective rules to distinguish strings of titles (e.g., "President, Secretary of State") from titles followed by names. In addition, we need more effective rules for distributing a title among all names of persons which follow it in the text (e.g., "Senators Vance Hartke and Birch Bayh of Indiana and Eugene J. McCarthy and Walter F. Mondale of Minnesota").

In addition, we may require rules which would *check* on the old ones rather than *supersede* them. The new set of rules would be applied to words and phrases which were identified as names of persons by the old set of rules. The new rules could indicate the degree of confidence with which the algorithm identified a word or a string of words as name of a person, or as a personal title followed by the name of a person, etc.

The advantage of this procedure consists in not having to revamp the algorithm in order to accomodate new rules. New rules would simply be tacked on to the old ones. New rules might check whether the elements of a string of letters, punctuation marks, spaces, numbers, etc. which the old rules had identified as a name can be (a) words of the English language and (b) names of persons.

Whether a string of characters is a personal name could be decided by probability tables constructed along these lines:

Is this string of letters an English word?	Can this string of letters be a personal name?	Is this string of letters the name of a person?
Yes	Yes	Probably yes
Yes	Unknown	It's unlikely
No	Yes	Yes
No	Unknown	It's very likely

The new rules should be relatively easy to implement. The question "Is this string of letters an English word?" could be answered by means of (a) a lookup in a dictionary based on some desk dictionary -- say Webster's Collegiate, and (b) simple rules for identifying affixes of the plural, the past tense, the gerund, the negation, etc. The question "Can this string of letters be a personal name?" could be answered by means of (a) a lookup in a dictionary based on a large telephone directory -- say the Manhattan Telephone Directory, and (b) simple rules for identifying the plural (e.g., "es" and "s" as in "the Joneses" and "the Weinbergs") and other affixes.

Improving the automatic identification system may require several subsidiary investigations. For instance, we may be well advised to determine the relationship -- if any -- between, on the one hand, the length, the date, the place of origin, the subject matter, the authorship, and the type of newspaper articles processed through the system, and on the other, the effectiveness of the algorithm.

Discussion and Interpretation.

Automatic classification of words and phrases of the type described here can be regarded as a particularly simple case of machine translation. However, the goal of this type of machine translation is not translation into another natural language but TEXT REDUCTION: certain words and word strings are identified as "pertinent" (e.g., personal titles, personal names, place names, street addresses, numbers and measures, dates and other time phrases, company names, trade names, chemical formulas, etc., etc.) and others as "not pertinent". Pertinent words and phrases are retained and labeled, and all others are suppressed.

Even this simple goal requires rules which are rather complex. However, because many word strings which the algorithms such as this one attempt to recognize have simple structure ("phrase structure"), they can be generated and possibly also recognized with a reasonable degree of accuracy by simple automata ("push-down storage") or by a combination of linguistic and statistical techniques.

More generally, it may be useful to view natural language as a macro-language containing certain special-purpose micro-languages (or "sublanguages") -- each with its own structure which relative to the total structure of language is quite simple. It may be of some practical and theoretical interest (a) to investigate the structures and the inter-relations of such sublanguages and (b) to construct algorithms for identifying in texts words and word strings belonging to such sublanguages.

An ability to produce and identify automatically words and word strings belonging to various special-purpose categories (i.e., sublanguages, each with its own set of rules) may prove to be very useful in information retrieval because they play an important role in various systems for extracting and distributing information.

It would appear that along with researching and developing methods for high-quality fully automatic classification of words in texts, it may be advisable to set up efficient procedures for (a) manual classification and tagging of words and word strings in texts, and (b) subsequent automatic extraction of data from texts which were recognized either manually or automatically. One procedure for manual classification of words in texts would require computer-legible texts which can be projected on TV-type tubes (hereafter, "display screens") and either lightpens or cursors for writing on display screens. It may look approximately as follows:

A newspaper article would be copied from some type of machine-readable tape into a suitable computer. The computer would then project the article on a display screen. A clerk would then scan the display screen and locate various types of words and phrases in the article (say, names of persons, names of organizations, dates, addresses, and so forth).

Upon identifying a type of word·or of word string, the clerk would flash a lightpen or a cursor at the display screen and bracket that word or word string in suitable identifying symbols. Next, identifying symbols would be transferred from display screen to tape by means of a computer program. The recognized tape could then be processed in various ways by miscellaneous information extracting programs.

It seems likely that manual assignment of word strings in texts to special-purpose sublanguages (akin to thesaurus classes) would provide a valuable interim service while methods for high-quality automatic classification are researched and developed. If and when automatic procedures for recognizing in texts dates, personal titles, various technical and professional terms, metalinguistic terms, names, etc., etc. become competitive with manual ones, the data processing community will be already in possession of operational computer programs capable of extracting data from recognized texts.

Some Possible Applications.

In the absence of figures on the cost of identifying personal titles and names by computer, the subject of the applications of computer programs capable of recognizing names of persons in newspaper texts must remain in the domain of speculations.

We would conjecture that if the speed of computation was high and its price could be kept low, and if the figure of merit could be raised to .98 or higher, then a computer program for identifying names of persons in texts would be worth incorporating into existing information retrieval systems of very large newspapers and periodicals.

It is still unknown whether a program with a figure of merit lower than .98 would be useful in information retrieval. We would surmise that it might be adequate for some purposes provided that it is sufficiently fast and cheap.

Several uses suggest themselves immediately for computer programs capable of identifying cheaply, rapidly, accurately, and exhaustively the names and titles of persons in computer-legible newspaper texts. They seem to fall into five broad and overlapping categories: (1) automatic indexing of newspaper articles, (2) determining how the names of persons cluster with one another and with other words, (3) establishing frequency counts of names of persons, (4) tracing associations between names of persons, and (5) answering questions of the "Who?" type within an automatic or semi-automatic system capable of providing answers to "Who?" "Whom?" "Whose?" "When?" and "Where?" types of questions addressed to a newspaper file.

Systems for (a) either automatic or manual classification of words and word strings in texts, and (b) subsequent automatic extraction of data from texts which were recognized either automatically or manually may be useful to many groups, among them: (1) political scientists, sociologists, lexicographers, onomasticians, and literary scholars concerned with the occurrence of names, titles, and other words in texts, (2) editors, documentalists, librairians, and others concerned with automation of editing and of literature searching, (3) opinion survey and market research statisticians concerned with the occurrence of names in texts, celebrity ratings, measurement of opinion trends, etc.

Footnotes.

[1] Borkowski, C.G., A System for Automatic Recognition of Personal Names in Newspaper Texts, Report RC-1563, Watson IBM Research Center, Yorktown Heights, N.Y., 1966, 62 pp.

[2] Meetham, A.R., Preliminary Studies for Machine Generated Vocabularies, Language and Speech, 6 (Part 1): 22-36 (January-March 1963).