Application of Automatic Language and Subject Identification for Universal Digital Library using Sparse Data

Lavanya Prahallad, Suryakanth V Gangashetty, Kishore Prahallad and Raj Reddy
School of Computer Science
Carnegie Mellon University (CMU)
Pittsburgh, PA 15213, USA
email: lavanyap@andrew.cmu.edu, svg@cs.cmu.edu, skishore@cs.cmu.edu, rr@cmu.edu

***Abstract*— In this paper we have proposed an approach for automatic language and subject identification for the books of digital library. The important characteristics of function words is explored for language identification. The heuristic search approach is explored for subject identification by matching title words with the keywords of the subjects. The language identification system is developed for five languages namely English, French, German, Italian and Spanish. The subject identification system is developed for books of English language. When the hypothesized language and subjects of a book are made as attribute of meta-data, then it will help Internet users find a book for a particular subject area of their interest in a language in an effective way.**

Index Terms **- Language Identification, Subject Identification, Sparse Data.**

I. Introduction

The primary objective of million book collection universal digital library (UDL) task is to capture million books in a digital format by scanning them at various scanning centers across the globe [1]. This task will also help the researchers who are working on improved scanning techniques, improved optical character recognition, and improved indexing approaches. The partners in the UDL project continue to digitize books at more than 50 scanning centers all over the world and are making steady progress towards the long-term objective of capturing all books in a digital format. One of the issue in UDL is how to digitize the books with high quality scanners and put them on-line in various formats user requires. A second might be to how to locate materials in the new digital library mainly the best search options and the simple yet powerful user interface. Yet another would be when to use and when to transcend the existing technologies and traditions of the physical library in its digital form. Still other issues stem from the problems of information overload created by new information technologies. However the main issue which needs to be addressed is the meta-data for the digital library books. But it is common to have simple and human errors while entering the meta-data. Its is important to have the accurate meta-data. We need to build the robust systems which automatically corrects the meta-data.

Many different meta-data formats exist. Some are quite simple in their description, others are quite complex and rich. When the language and subjects of a book are attribute of meta-data, then it will help Internet users find a book for a particular subject area of their interest in a language in an effective way. This paper focuses on automatic identification of the language and subject to which the books belong by inspecting titles for million book collection project. The language and subject identification helps in creating the meta-data for the scanned books. It is well known that languages rarely borrow function words from other languages or make up new ones [2]. In this paper we have proposed an approach for automatic language identification by exploring the important characteristics of function words. Using heuristic search approach and stored knowledge, we propose identification of subject area to which a title of the book belongs.

This papers is organized as follows: Section II describes the necessity of automatic language and subject identification for books of universal digital libraries. Section III describes the approach and studies on automatic language identification. Section IV describes the approach and studies on automatic subject identification. The final section summarizes the studies.

II. Necessity of Language Identification for Books in Universal Digital Libraries

In the digital libraries different language books will be scanned and their meta-data will be entered in the local languages for the convenience for the people to read in their local languages. Suppose when they are written in English, we can read the title, but we cannot figure out to which language the book belongs. When it is written in a local language, one cannot make out which language the book belongs to as we cannot know all the languages unless we see the administrative meta-data. So the language identification tool will be useful to find out which language the book belongs to, even though the meta-data is not entered for the book at the time of scanning. Otherwise we have to just depend on the data entry operator to enter the language of that book in the language field. Since it is done by humans, there is a possibility of entering wrong

language for a book. The similar argument is applicable for subject area of the book.

The automatic language and subject identification process ensures the accuracy in finding out of the language and subjects of the books. Presently we have developed an approach for automatic language identification for the five languages namely English, French, German, Italian and Spanish. Further we have developed an approach for automatic subject identification for the English language books. In the next section we describe the studies on automatic language identification.

III. Automatic Language Identification

Words are divided into two categories namely function words and content words [3]. Linguists usually draw a distinction between content words, those words whose meaning is best described in a dictionary and which belong in open sets so that new ones can freely be added to the language, and function words, words with little inherent meaning but with important roles in the grammar of a language. Function words are closed class words while content words are open class words. Function words (or grammatical words) are words that have little lexical meaning or have ambiguous meaning [4]. But instead serve to express grammatical relationships with other words within a sentence. Function words may be prepositions, pronouns, auxiliary verbs, conjunctions, grammatical articles or particles, all of which belong to the group of closed class words [5]. Since function words belong to the closed class of words in grammar in that it is very uncommon to have new function words created in the course of speech, whereas in the open class word, that is nouns, verbs, adjectives, or adverbs, new words may be formed readily [6] [7]. We explore these important properties of function words for automatic language identification task.

A. Approach for Automatic Language Identification

The following are the proposed steps for automatic language identification.

1) Select the most frequently occurring function words in each language
2) Create dictionary of function words for each language
3) Extract the words from the title of a book
4) Get the statistics of number of words matches with number of function words (stored in dictionary) in each language
5) Hypothesized language of the book is the one whose title words matches with maximum number of function words of that language.

The block diagram of the proposed system for automatic language identification is shown in Figure 1.

B. Studies on Automatic Language Identification

In this studies we have considered the UDL database for language identification. It has records for 1535764 books. Each record has bar-code and title of the corresponding book. The Table I gives a typical sample entries in the database file. The corresponding target languages are, English, French, German, Italian and Spanish respectively. Out of 1535764 books, there are 365518 books belongs to these five languages. In Table II we have given the sample list of function words considered for each language. The confusion matrix obtained by using automatic language identification approach is given in Table III. Table IV summarizes the results of language identification. It is seen that the approach has provided more than 94% of the time the correct language identification in all the five languages. In the next section, we describe the studies on automatic subject identification.

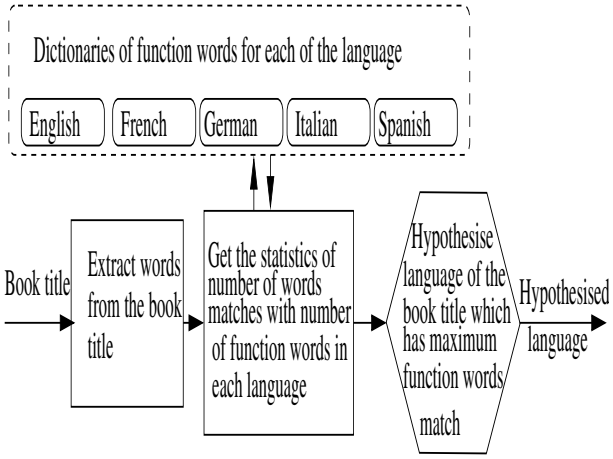

Fig. 1. Block diagram of the system for automatic language identification.

TABLE I

Sample entries in the database file.

Bar-code	Book Title
651835	Lost Train Over Rostov Bridge
651846	La Literatura Espanola Resumen De Historia Critica
770154	Der Stammbaum Der Inflecten
659963	I Canti Con La Vita Del Poeta
646196	El Capitan Veneno

TABLE II

List of function words.

Language	Function Words	
	Number	Sample List
English	98	a he in is it of that the to was all are as at be but for had have him his not on one said so they
French	84	de la le et les des en un du une que est pour qui dans a par plus pas au sur ne
German	101	der die das und sein in ein zu haben ich werden sie von nicht mit es sich auch auf f an er so
Italian	100	non di che e la un il a per in una mi sono i ho l lo le ma ha ti si con se da come ci cosa io
Spanish	90	de que no a la el es y en lo un por me qu1 te los se con para mi est si pero las su yo tu del al como

IV. Automatic Subject Identification

In this section we describe our proposed approach based on heuristic search and stored knowledge to identify subject area of the book title.

TABLE III

CONFUSION MATRIX FOR DIFFERENT LANGUAGES. THE NON-DIAGONAL ENTRIES IN THE TABLE FROM COLUMNS 2 TO 6 REPRESENT THE MISCLASSIFICATION. THE DIAGONAL ENTRIES REPRESENT THE CORRECT CLASSIFICATION.

Languages	English	French	German	Italian	Spanish
English	357298	1125	2428	1429	228
French	4	741	6	12	20
German	5	14	423	4	1
Italian	1	6	1	1681	1
Spanish	2	0	0	2	86

TABLE IV

RESULTS OF LANGUAGE IDENTIFICATION.

Language	Number of Titles		
	Used	Correct Identification	%Correct Identification
English	362508	357298	98.56
French	783	741	94.63
German	447	423	94.63
Italian	1690	1681	99.46
Spanish	90	86	95.55

A. Approach for Automatic Subject Identification

The steps involved in identifying the subject from the title is as follows:

1) Consider only the first 8 characters as stems for subject names and the the corresponding keywords of each subject.
2) Similarly consider only the first 8 characters as stems of each word in the title
3) Match the title stems with the patterns of each subject (The sample list of standard patterns for subjects is given in Table V). If no subject is found proceed to step 4.
4) Match the title stems with the stems of the subject names. If no subject is found proceed to step 5.
5) Match the title stems with the stems of the keywords of each subject. If no subject found proceed to step 6.
6) Match the title stems with the stems of subject "General". If no subject found, then consider the title as undetermined.

In steps 3, 4, and 5, often it is possible the title matches with more than one subject. To handle these cases, we maintain a hand-coded hierarchy of the subjects. For example, if a title matches with Electrical Engineering and Physics, a preference is given to specialized subject such as Electrical Engineering as apposed to Physics.

B. Studies on Automatic Subject Identification

In this studies we have considered the UDL database for subject identification of English books. It has records for 1535764 books. Each record has bar-code and title of the corresponding book. The first two columns of the Table VI gives the typical sample entries in the database file. The third column gives the corresponding hypothesized subject. Out of 1535764 books, there are 358401 books belongs to English language. The Table VII gives the list of 48 subjects considered in our studies. The Table VIII gives the results of automatic subject identification by our proposed approach. It is seen from the results that about 63% of the books are correctly identified their subject areas without ambiguity.

TABLE V

SAMPLE LIST OF STANDARD PATTERNS FOR SUBJECTS IN ENGLISH LANGUAGE.

Standard patterns	Hypothesized subject
civil war	History
foreign exchange	Economics
neural network	Artificial_intelligence
live saint	Religion
town life	Sociology
infra red	Physics
human nature	Psychology
north pole	Geography
memoirs of	Biography
weather forecast	Meteorology

TABLE VI

SAMPLE ENTRIES IN THE DATABASE FILE.

Barcode	Book Title	Hypothesized subject
644130	Impressions of Latin America	History
644185	The physics of music	Music
649808	The bible in Spain	Religion
650428	The book of the ancient Romans	Literature
652268	The manual training school	Education
661127	Characteristically American	Drama
675820	Planetary and space science	Astronomy
673726	People's democracies	Politics
673739	This business of exploring	Marketing
814539	Analytic syntax Literature	Language

TABLE VII

LIST OF (48) SUBJECTS CONSIDERED FOR ENGLISH LANGUAGE.

List of Subjects
Agriculture, Anthropology, Architecture, Art, Artificial_intelligence, Astrology Astronomy, Biography, Biology, Biotechnology, Chemistry, Civil_engineering, Computational_science, Culinary, Defense, Drama, Economics, Education, Electrical_engineering, Environmental_engineering, General, Geography, History, Language, Law, Literature, Marketing, Material_science, Mathematics, Mechanical_engineering, Medical, Meteorology, Music, Mythology, Organizational_behavior, Philosophy, Physics, Poetry, Politics, Psychology, Religion, Robotics, Sociology, Software, Sports, Statistics, Technology, Zoology,

V. SUMMARY AND CONCLUSIONS

In this paper we propose automatic language and subject identification of the books. This will be useful for the meta-data in universal digital library. Through the proposed approaches and series of experiments, the performance of our approach shows the comparable power to that of human readers. Still it is necessary to to achieve better performance in subject identification. Basically, automatic language and subject identification would reduce the human efforts that are involved in library. The language and subjects are possibly two important meta-data to describe to which language and subject a book belongs. This helps to search for specific list of books by considering the subject and language as a search option.

TABLE VIII

RESULTS OF SUBJECT IDENTIFICATION.

Titles found a match with	Number of Titles	Results Statistics
Single subject	229047	63.90% Correct Identification
More than one subject	40538	11.31% Ambiguous
No subject	88816	24.79% Needs further processing

However from the view point of universal library development, it is necessary to provide numerous meta-data for readers or users. We continue to work on automatic subject and language identification task for other languages as well.

REFERENCES

[1] Source, http://www.ulib.org/, ," .

[2] Gelderen Elly van, "The Rise of Functional Categories," *Amsterdam and Philadelphia: Benjamins.*, 1993.

[3] Abney Steven, *The English Noun Phrase in Its Sentential Aspect*, PhD thesis, Massachusetts Institute of Technology, 1987.

[4] Lightfoot, David, *Principles of Diachronic Syntax*, Cambridge and New York:, Cambridge University Press, 1979.

[5] Traugott Elizabeth and Berndt Heine, "editors, Approaches to Grammaticalization, 2 vols.," *Amsterdam and Philadelphia: Benjamins.*, 1991.

[6] Radford Andrew, *Syntactic Theory and the Acquisition of English Syntax*, Oxford and Cambridge, Massachusetts: Blackwell, Blackwell, 1990.

[7] Radford Andrew, *Syntactic Theory and the Structure of English*, Cambridge and New York:, Cambridge University Press., 1997.