AUTOMATIC LANGUAGE IDENTIFICATION AND TRANSLATION

Sayali Nandanwar, Ravina Jain, Prajakta Kuralkar, And Rasika Sambare

Department of Computer Technology,

KDK College of Engineering, Nagpur

Abstract—This paper deals with the automatic identification of natural languages and their translation .We intend to use an alternative approach to the standard use of both K-means and Ant Class Clustering algorithms. We propose to use a multilingual text corpus to assess this approach. Taking into account that this method does not require a priori and information (number of classes, initial partition), is able to quickly process large amount of data, and that the results can also be visualised.

Index Terms—K-means , Language identification , clustering , multilingual texts , Ant-Class

INTRODUCTION

With the advent of technology, the world is growing smaller by the day, but the number of languages known to the people has remained the same. Hence, there is need of a system which can break these lingual barriers and help people communicate better wherever they go. This project deals with the same. It has various features that are the perfect package for a user to understand any audio or a visual in his/her desired language.

This is speaker independent. Also a user can know the accuracy rate of translation and learn various languages . Research in recent years has given a lot of interest to textual data processing and especially to multilingual textual data. This is for several reasons: a growing collection of networked and universally distributed data, the development of communication infrastructure and the Internet, the increase in the number of people connected to the

global network and whose mother tongue is not English . This has created a need to organize and process huge volumes of data. The manual processing of these data (expert knowledge based systems) is very costly in time and personnel, they are inflexible and generalization to other areas are virtually impossible, so we try to develop automatic methods .

One of the major issues raised in any application of automatic processing of digital documents is that of multilingualism, since we want to perform linguistic processing. Any linguistic text processing is completely dependent on the language of the latter. It is therefore essential in a multilingual environment that research tools are able to automatically identify the

languages of the documents they have to deal with. A variety of methods for identifying text language of a multilingual corpus have been developed. We propose in this paper to study the effectiveness of the clustering algorithm Ant Class for identifying text language of a given multilingual corpus, based on a vector representation focused not on words but on the n-grams for representing the texts.

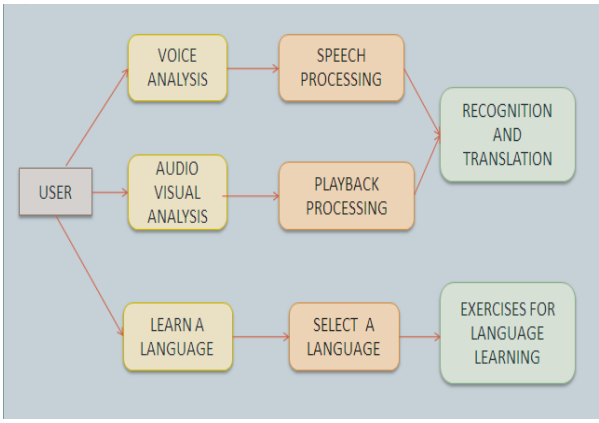
I. PROBLEM DEFINITION:

Automatic Language Identification and Translation :

Automatic and spontaneous analysis and recognition of the language to be translated from , in real-time constraint , for :

- audio recordings,
  and
- playback in an audio-visual.

II. DATA FLOW DIAGRA



III. AUTOMATIC LANGUAGE IDENTIFICATION

Language identification is to assign a textual unit, supposedly monolingual to a language. This identification became important as textual data in different languages, are more and more available on the global network . Automatic language identification is possible because natural languages are extremely non-random, and they each have regularities in the use of characters or character sequences.

The alphabet of each language is either unique or highly characteristic of this language. Information on the stability and consistency of the frequency of letters and letter sequences are not new . It is statistically proven that for each language, the number of occurrences of the sequence of two, three, four or five letters are stable and different from language to language. For example, in English, in any text, the frequency of the letter "E" is about 13%, the frequency of the letter "U" is about 3% and the frequency of the letter "Z" is approximately 0.1%. For two sequences of characters or bi-grams, we find for example that the probability of having the string "TH" in English is relatively high, in Spanish and Portuguese, this probability approaches zero. In the same order, the probability of having "SZ" in Hungarian and Polish is great; the string "TION" characterizes the French and English. Based on these probabilities of occurrence of letters and letter sequences, we can design an algorithm capable of identifying the language of a text.

We can distinguish two kinds of approaches: linguistic approaches, and

probabilistic and statistical approaches.The approaches based on linguistic knowledge involve the construction of linguistic resources and require prior knowledge. They are not generalizable to the classification of

languages in text categorization.The statistical approaches use probabilities and knowledge built automatically from a

text corpus representative of the language, the goal is to capture using statistical models and probabilities certain regularities of the languages and their associated frequency orprobability of occurrence.

They generalize the recognition of language classification of texts. These empirical regularities play the role of linguistic knowledge. The identification is to calculate the probability for a statement to belong to different languages, according to the regularities observed.

A variety of tools have been developed to classify texts based on their respective languages. However, all these approaches work in a supervised manner given a sample of each language model, parameters are estimated for prediction and texts are classified according to their similarity with the learning text sets. But supervised learning has a major drawback: The languages that are not contained in the training set will not be identified and the text will be assigned to other classes arbitrarily.

We show in this report , a method that operates on the n-grams of characters as attributes, and clusters together similar texts and discovers the number of languages in a

completely unsupervised manner.

VI . METHODOLOGICAL APPROACH

In this section we describe our methodology. With the use of the approach based on the n-grams we construct matrix documents-terms that will be exploited by the Ant Class algorithm to group similar documents together. This combination will be examined in several experiments using the Euclidean distance, cosine distance and Manhattan distance as similarity measures for several values of n.

A. Corpus

The texts of our corpus come in several formats (HTML, XML and SGML). For each item of the corpus, we wish to remove all tags like: <title> ... </ title><author> ... </ author><date> ... </ date> ... etc... and take only the text part (written by the author).

We shall follow the below described approach :

In a first step, we will transform the uppercase characters to lowercase characters for English and French, then we eliminate from the text diacritical

characters (punctuation) such as: dot, comma, semicolon, the question mark and exclamation etc.... and the numbers because these characters tend to have no influence on the

results of the clustering and do not provide relevant information for the decision making, their elimination reduces the size of the representation space. The corpus texts

will be saved in UTF-8 encoding. This will allow us to handle documents that use different character sets.

B.A Representation Based On n-gram of characters

The term "n-gram" was introduced in 1948. Since then, the n-grams have been used in several areas, such as speech recognition systems, with typical values of n equal

to 3 or 4. They are now also used in systems for automatic processing of language for information retrieval. One of the applications of the n-grams model is the indexing of

large corpus . An n-gram may designate both an n-tuple of characters (n-gram character) or an n tuple of words (n-gram words). This model does not represent documents by a vector of term's frequencies, but by a vector of n-gram's frequencies in the documents.

An n-gram character is a sequence of *n* consecutive characters. For any document, all n-grams that can be generated are the result obtained by moving a window of *n* boxes in the text. This movement is made in stages; one stage corresponds to one character for n-grams of characters, and a word for n-grams of words. Then we count thefrequencies of n-grams found. In scientific literature, this term sometimes refers to sequences that are neither ordered nor straight, for example a bigram can be composed ofthe first letter and third letter of a word; consider an n-gram as a set of unordered *n* words after performing the stemming and the removing of Stopwords.Techniques based on n-grams have several advantages: they automatically capture the roots of the most frequent words and operate independently of languages and are tolerant of spelling errors and distortions caused when using optical scanners and do not need the removing of Stopwords or the stemming process [that improve the performance of words based systems. An n-gram refers to a string of *n* consecutive characters.

In this approach, we do not need to conduct a linguistic processing of the corpus. Fora given document, as we already said, extracting all n-grams (usually $n = (2, 3, 4, 5)$ is the result obtained by moving a window of *n* boxes in the main text. This movement is made by steps of one character at a time, every step we take a "snapshot" and all these 'shots' constitute the set of all n-grams of document.We cut the texts of the corpus based on the value of *n* chosen. We took $n = 2, 3, 4$and 5.

For example, the 5-grams characters of the following text :

Sandoz ag said it planned a joint venture to produce herbicides in the soviet union the company said it had signed a letter of intent with the soviet ministry of fertiliser production to form the first foreign joint venture the ministry had undertaken since the soviet union allowed western firms to enter into joint

ventures two months ago the ministry and sandoz will each have a stake but a company spokeswoman was unable to give details of the size of investment or planned output

"Sando, andoz, ndoz_, doz_a, oz_ag, z_ag, _ag_s, ag_sa, g_sai, _said, said_, aid_i, id_it, d_it_, _it_p, it_pl, t_pla, _sitemap ..., ned_o, ed_ou, d_out, _outp, Outpu, utput" The character _ represents a space.

We constitute in this way the cross table *Nij* of occurrences of the n-gram i in text *j* so

that all the n-gram do not contain spaces and belong absolutely to an index of Arabic, English, French, Spanish and Italian , predefined in advance.

Algorithm n-gram

(1) for each text do

(2) for each n-gram do

(3) if the n-gram contains a "" (Space) then remove the n-gram

(4) else if the n-gram belongs to the index then check for an entry in the

global vector of the n-grams, which corresponds to this n-gram,

increment the box *Nij* where *i* is the rank of the n-gram in the

global vector and *j* is the text number

(5) else create a new entry corresponding to this n-gram in the global

vector and affect 1 to *Nij* where *i* corresponds to the last n-gram

and*j*to the text number

(6) end if

(7) end if

(8) end for

(9) end for

C. Dimension Reduction

The objective of reduction methods of terms is to provide a shorter but more meaningful list of terms. The terms are usually ordered from the most important to least important according to some criterion. The question arises in the number of terms to retain in the list. To choose the right number of words, you must know whether the information conveyed by the words at the end of the list is useful, or it is redundant with information provided by the terms of the beginning of the list. There is no evidence that a large number of terms is necessary for good performance, because even with models like Support Vector Machines (SVM) which are in principle suitable for large vectors, the results are contradictory. This is probably due to the fact that the terms are mutually correlated, and to the way different algorithms manage these relationships. We know that reducing the size by using the frequency-document is immediate, and that its performance is equivalent to other more sophisticated forms despite its simplicity. It eliminates the n-grams that appear in a number of documents below a certain threshold. We chose to eliminate the n-grams that appear in only one document (the chosen threshold is 1), greatly reducing the number of n-grams.

At the end of these steps, we obtain a document-term matrix Nij and an overall n-grams vector (n = 2, 3, 4 and 5).

To calculate the weight (frequency) of each extracted n-gram, we use a combination of local and global weights.

D. Clustering multilingual texts

Several clustering methods have been applied to textual documents . The Ants which possess a range of behaviors very diverse (collective or individual) suggest very interesting heuristics for many problems including clustering. An early study on this area was conducted where a population of ant-agents moves randomly on a two dimensional grid and are able to move objects in order to gather them. This method was extended by on simple objects. An extension of the algorithm "LF" of was

presented when researcher's developed an algorithm called Ant Class using the same principles that LF and adding some improvements.

In LF each cell can contain only one object, a class is then represented by a cluster of objects. In AntClass several objects can be placed on a single cell (the ants can pile up objects in the same grid cell), forming a pile. In this case, a class corresponds to a pile and a partition is given by all present piles in the grid. Each pile has a representative which is the center of gravity of the elements that constitute it. This is a hybrid with the K-Means algorithm.

This hybridization consists in initializing the K-Means algorithm with the partition obtained by grouping objects ants. Thus, this new principle allows for automatic interpretation of classes which is done visually and with more difficulty in LF. Moreover the Ant Class algorithm converges faster, as in LF an ant can pass a number of iterations to find an empty slot next to the group of objects close to that it carries.

The grid G is square and its size is determined automatically based on the number of objects to be treated. If N is the number of objects, G contains L cells per side:

L =|sqrt(2N)|, this formula ensures that the number of cases is at least equal to the number of objects.

Initially the A ants {a1,. . ., aA}

are arranged randomly on the grid by checking

that a cell can only accommodate a single ant and come with a carrying capacity c(ai), a memory of size m(ai), velocity v(ai)and a patience p(ai), knowing that T is the number of moves of each ant.

Algorithm Ants: grouping objects by ants.

Ants(Grid G)

(1) for t= 1 to T do

(2) for k= 1 to Ado

(3) Move the ant akon one cell unoccupied by another ant

(4) if there is a lot of objects Tj on the same cell that akthen

(5) if the ant ak is carrying an object oi[a lot of objects Ti] then

(6) place the object oi

[the pile Ti] carried by the ant on the pile Tj

following the probability

pd (oi, Tj)[pd(Ti, Tj)]

(7) else Pick up the object oithe most dissimilar of the pile Tj[until the capacity c(ak)of the ant is reached or the pile is empty] by the probability pp(Tj)

(8) endif

(9) endif

(10) endfor

(11) endfor return the grid G

E. Language Identification

In this step, we assign a label to each class (after the clustering process) matching the dominant language of each pile of the grid.

We proceeded as follows:

- For each pile, we specify the n-grams that appear at least once in the text of this pile.

- We calculate for each pile the percentages of its component languages as follows:

-For each gram we traverse the Arabic, English, French, Spanish and Italian index predefined in advance and point the language in which the gram appears by incrementing a corresponding counter;

-For each language we divide the corresponding counter on the total number of terms of this pile;

-At the end of this stage, we determine the rate of text for each language present in the pile.

- For each pile, we assign a label named after the dominant language in this pile.

- If we find piles of the same labels, we merge them into a single class, to obtain the minimum number of classes with distinct labels.

Thus we obtain the classes of languages.

VII. FUNCTIONS:

This project consists of three core modules which have the following primary functions :-

1. Recording the voice of a person, analyzing it , and providing the translation of that speech in a user desired language.

2. To provide translation of the playback in a video into user desired language in the form of an audio.

3. This project also gives an add-on to the user to learn various languages in their basics.

4. Providing user with the translation accuracy percentage.

VIII.CONCLUSION:

Thus , we conclude , that -

- This application will provide a platform for a user to analyze voices and audio -visuals.

- Along with analyzing , it also automatically identifies the language of text as well as speech and then , translates that language into user desired language .

- Also, it provides the basic features to learn different languages.

Moreover, the method illustrated in this paper, shows that it is possible to identify automatically the language in an unsupervised manner and, aims to enhance the unsupervised methods and

techniques applied to classification for text language identification of a multilingual corpus.

We gave a method based on the behavior of real ants having collective and individual characteristics and ability to gather and sort objects. The Ant Class algorithm shown is hybrid; the search of the number of classes is performed by the artificial ants algorithm and a conventional classification algorithm the K-means, is used to correct the misclassification inherent to stochastic

method such as artificial ants. This method is also characterized by the fact that it does not require a priori information (number of classes, initial partition), and possibly even no parameters, and is able to quickly process a large amount of data. The results provided by our methods can also be visualised. We realized that the choice of a similarity measure is crucial in the process of clustering. Indeed, two different measures can lead to two different results of clustering.

IX. REFERENCES:

*Abdelmalik Amine , Zakaria Elberrichi ,Michel Somonet. Automatic language identification : An alternative unsupervised approach using a new hybrid algorithm

*Peters C, Sheridan P. Multilingual Information

Access. In M. Agosti, F. Crestani, G. Pasi

(eds.). Lectures on Information Retrieval,

Lecture Notes in Computer Science

1980, pp51-80, Springer Verlag, 2001.

*Sebastiani F. Machine learning in automated text categorization.ACM Computing Surveys, 2002, 34(1): 1–47.

*Hughes B, Baldwin T, Bird S,Nicholson J, and MacKinlay A. Reconsidering language identification for written language resources. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), 485–488, 2006, Genoa, Italy.

*Monmarché N, Slimane M, Venturini G. On improving clustering in numerical databases with artificial ants.In D. Floreano, J.D. Nicoud, etF.Mondala, editors, 5th European Conference on Artificial Life (ECAL'99), Lecture Notes in Artificial Intelligence,volume 1674, pages

626–635, Swiss Federal InstituteofTechnology,Lausanne, Switzerland, 13-17 September 1999. Springer-Verlag.

*Řehûrek R, Kolkus M. Language

Identification on the Web: Extending the Dictionary

Method.In Computational Linguistics and Intelligent Text Processing, 10th International Conference, CICLing 2009, Proceedings.Vyd.první. Mexico City, Mexico: Springer-Verlag, 2009. ISBN 978 -3 -642 -00381 -3, pp. 357 -368.

*Beesley K. Language Identifier: A Computer Program for Automatic Natural Language Identification on On-Line Text.In Proceedings of the 29th Annual Conference of the American Translators Association, 1988, pages 47– 54.

*Cavnar W B, Trenkle J M. N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval, 1994, pages 161–175, Las Vegas, US.

*Dunning T. Statistical Identification of Languages. Technical Report MCCS, 1994, 94-273, Computing Research Laboratory.

*Schütze H, Hull D A, Pedersen J O. A comparison of classifiers and document representations for the routing problem. In Fox, E. A., Ingwersen, P., and Fidel, R., editors, Proceedings of SIGIR-95, 18th ACM International Conference on Research and Development in Information Retrieval, 1995, pages 229–237, Seattle,US. ACM Press, New York, US.

*Shannon C. The Mathematical Theory of Communication. Bell System Technical Journal, 1948, 27: 379–423 and 623–656.

.