Paper 096-2012

Translating Foreign Language in SAS® with Google Translate

Murphy Choy, School of Information System, SMU, Singapore

ABSTRACT

Increased levels of global operation have resulted in companies having business in many different geographical regions of the world. This move necessitates the operation to be relatively localized and, thus, results in databases with extensive amounts of information to be captured in a foreign language. Often this creates difficulties in analytic projects that are done by English-speaking analysts operating in non-English-speaking environments. With the advent of Google Translate services, we will demonstrate how a simple command issued to Google can be used within SAS to generate the translated results.

INTRODUCTION

Managing analytics project is always a tussle for time. The race to complete any analytic project is usually hampered by the exceeding long amount of time doing data management and cleaning. This process is usually the most important part of any analytics project and any mistake at this stage will return to haunt the rest of the project and can be extremely costly to the project sponsor. This process is further fraught with difficulties when the project is managed in a country where most of the data are stored in languages other than English. This problem is encountered commonly in Asian Countries such as Thailand, South Korea, Taiwan, Japan and China. These Asian countries typically operate with their national language which is not English. For analysts operating in an offshore environment, this is extremely undesirable as the data may be transferred to them in the native language making data understanding impossible. Hiring a conversant analyst with the local language is often a desirable option but hardly practical given the diverse amount of languages in use in many countries. One of the simplest way is to parse the data into a translator and then passing it back into SAS. There are many translation software available and in this paper, we will focus on the simplest approach of using Google Translate.

GOOGLE TRANSLATE

Google translate is a free translation service provided by Google on their website. Google translate is a statistical machine translator that was based on the design proposed by Franz Josef Orh. Traditional machine translation is usually rule based in nature. However, there are some serious limitations of the approach as it strictly follows the prescribed grammatical rules which reduce the ability of the machine to adapt to new developments in the language. At the same time, the ability of the machine to adapt to certain level of localization of language will be extremely difficult and translation rules will be hampering the performance of translation. To achieve a high level of precision in terms of words translated, a huge database is needed which is usually populated by documents translated by professional translator between the two language involved. Google translate uses the translated documents between multiple languages in the United Nations as the foundation for the translation services resulting in a gigantic corpus to rely on. However, this does not eliminate all errors with the translation. There are still some forms of translational error that will occur in certain situation. Some of the more commonly encountered errors include similar but non-equivalent substitution and inversion of sentences which is common for Asian languages.

CHOICE OF TRANSLATION SOFTWARE

With the presence of translation errors, many people would question the choice of using Google translate as the platform for translating documents. There are some important aspects of the data which will drive the choice of the software. First and foremost, unlike traditional translation needs in area of speeches, translations in data base tend to focus on simple phrases in the local language which is easily accomplished with translators. At the same time, Google translate is easily available to users with internet access and is free-of-charge with API available to users. These few factors affected the decision to use Google translate.

USING SAS TO PARSE DATA TO GOOGLE TRANSLATE

To initiate a translation in Google translate, one can use the url address of Google translate to send the words to translate. Let us use the example of a French phrase *Bonjour Madame* in which we seek to translate it from French to English. Below is the example syntax to achieve this.

```
http://www.google.com/dictionary?langpair=fr|en&q=bonjour+madame&hl=en&aq=f
```

Note the particular syntax in the URL. The language pair is declared in the URL. Below is another example from English to German for the phrase *My Lady*.

```
http://www.google.com/dictionary?langpair=en|de&q=my+lady&hl=en
```

Thus we can see that Google accepts our query easily via the use of a URL statement. However, how do we extract the information needed? The key to this lies in the URL option in filename. The FILENAME statement in SAS is often used to point to URLs which are then processed by SAS into SAS dataset. However, in most data, there are a number of roles which requires cleaning. To achieve this, we will be relying on the nested programming logic in SAS. This process is aided by the Call Execute function in base SAS which can run codes with input from the data containing information. Once the call has been made, a data set is created for each word and then appended and merged together to form the data containing the translated text. Combining these qualities together, we developed a simple macro in Appendix A to achieve the translation from different languages to English.

CONCLUSION

Translation between languages in SAS data can be easily achieved with the application of a simple macro and Google translate. This approach also allows users to do translation whenever they have internet connection.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Name: Murphy Choy
Enterprise: School of Information Systems, Singapore Management University
Address: 80 Stamford Road
City, State ZIP: Singapore 178902
Work Phone: +65-92384058
E-mail: goladin@gmail.com/murphychoy@smu.edu.sg

APPENDIX A

```
/*************************************************************************
GOOGLE TRANSLATE MACRO
*************************************************************************/
%MACRO RESULTEXTRACT(INPUT,WORDS,OUTPUT);

/*FILENAME DECLARATION STEP*/
DATA _NULL_;
SET &INPUT;
LENGTH G1 $55.G2 $255.;

G1 = TRANSLATE(&WORDS,'+',' ');/*REMOVE SPACES*/

/*CHANGE IF LANGUAGE IS SOMETHING OTHER THAN FRENCH TO ENGLISH*/
G2 = "FILENAME SCB"||COMPRESS(_N_)||" URL
"||COMPRESS("'http://www.google.com/dictionary?langpair=fr|en&q="||G1||"&hl=en&aq=f'")||" lrecl =
32000;";/*CONSTRUCT CALL TO GOOGLE TRANSLATE*/
CALL EXECUTE(G2);

CALL SYMPUT('TOTNUM',_N_);/*INDICATE NUMBER OF STATEMENTS*/

RUN;

/*************************************************************************/
/*LOOP TO CALL THE TRANSLATION*/
%DO I = 1 %TO &TOTNUM;

        /*EXTRACTING THE TRANSLATION*/
        DATA TEMP;
        INFILE SCB&I LRECL = 32000 TRUNCOVER;
        INPUT A $ 1-4096;
                TRANSLATEDWORDS = A;
                IF ANYPUNCT(TRANSLATEDWORDS,1) = 0 AND NOT MISSING(TRANSLATEDWORDS) THEN OUTPUT;
        RUN;

        /*APPENDING THE RESULTS*/
        PROC APPEND BASE = TRANSLATEDWORDS DATA = TEMP;RUN;

        %END;

/*CREATE THE TRANSLATION TABLE*/

DATA &OUTPUT;
```

```
MERGE &INPUT TRANSLATEDWORDS(KEEP = TRANSLATEDWORDS);
RUN;

%MEND;

/*************************************************************************/

/*EXAMPLE*/

DATA TTT;
        WORD = "bonjour madame";OUTPUT;
        WORD = "jeanne de arc";OUTPUT;
        WORD = "femme fatale";OUTPUT;
        WORD = "femme séduisante";OUTPUT;
RUN;

%RESULTEXTRACT(TTT,WORD,TTT2);

/*************************************************************************/
```