

English to Creole and Creole to English Rule Based Machine Translation System

Sameerchand Pudaruth, Lallesh Sookun, Arvind Kumar Ruchpaul
Computer Science and Engineering
University of Mauritius
Mauritius

Abstract— Machine translation is the process of translating a text from one language to another, using computer software. A translation system is important to overcome language barriers, and help people communicate between different parts of the world. Most of popular online translation system caters only for the most commonly used languages but no research has been made so far concerning the translation of the Mauritian Creole language. In this paper, we present the first machine translation (MT) system that translates English sentences to Mauritian Creole language and vice-versa. The system uses the rule based machine translation approach to perform translation. It takes as input sentences in the source language either in English or Creole and outputs the translation of the sentences in the target language. The results show that the system can provide translation of acceptable quality. This system can potentially benefit many categories of people, since it allow them to perform their translation quickly and with ease.

Keywords— rule-based; Mauritian Creole; English

I. INTRODUCTION

People around the world have access to millions of documents, articles and websites over the Internet. However these electronic documents are often written in different languages and therefore it is necessary to translate them into the relevant target language.

Traditionally, human translators have assisted people to understand texts in a foreign language. However the translation process is very time consuming and costly, when it is done by a human translator. Therefore there is an increasing need for a translation tool so that communication can take place between different parts of the world. A translation tool would help to overcome language barriers.

In Mauritius, even if English is the official language, the majority of the Mauritian population uses the Creole language as a medium of communication. It is a language which is spoken only in Mauritius and some other small islands in the Indian Ocean. Many Mauritians face difficulties in expressing themselves properly using the English language. At the same time, foreigners and most particularly tourists find it extremely difficult to communicate with the Mauritian people.

This is mainly because many Mauritians are at ease only in their native language, which is Creole. The Mauritian Creole language has recently been introduced in the education system in Mauritius [1] as an optional subject and also as a medium of instruction to facilitate teaching and learning in primary schools. The introduction of Creole at school is due to many

factors, the most important of which is the high rate of failure at the Certificate of Primary Education (CPE) exams.

Mauritian Creole is thus becoming more and more important for the Mauritian population as it has already been formalized as a full-fledged language. Its usage in formal situations has increased dramatically over the last five years. There was a time when it was considered rude to do advertising in Creole. However, now things have changed considerably and Creole has become the preferred medium for advertising for all range and types of organisations as well as for the government. Even the Bible is now available in Creole.

In this paper we attempt to solve the identified problems by introducing a machine translation system that will help people translate texts from English to Mauritian Creole and vice-versa. The translation system would greatly assist students in switching to and from Mauritian Creole and English language. It would also be vital to foreigners and tourists as it would enable them understand the meaning of certain words or texts in the Creole language.

The tool developed can also be used in conjunction with other translation tools. Thus, A German can use Google translate to translate German texts into English and the use our tool to convert to Creole and vice-versa. Since the Mauritian economy depends heavily on Tourism and tourists from all over the world come to Mauritius, the tool can be of tremendous value to visitors.

This paper is organized as follows. Section 2 looks at the related work. The section is divided into 4 parts: an introduction to machine translation, its architectures and briefly describes the main paradigms that have been developed. An overview of the Mauritian Creole grammar will then be discussed. In section 3, we present how the translation system has been implemented. Finally, we evaluate the system in Section 4 and conclude the paper in Section 5.

II. RELATED WORKS

There are a number of issues that needs to be address for rule-based machine translation systems. For example, Charoenpornawat et al. [2] address the issue of word-sense disambiguation by using machine learning techniques to automatically extract context information from a training corpus.

Their work improves the translation quality of rule based MT systems using this approach. Oliveira et al. [3] shows that by using a systematic approach to break down the length of

the sentences based on patterns, clauses, conjunctions, and punctuation can help to parse, and translate long sentences efficiently. Also, Poornima et al. [4] suggest simplifying complex sentences into simpler sentences in order to improve the translation quality. Their research presented results which showed that this method was able to preserve the meaning of the sentence after translation.

A. Machine translation

Machine translation (MT) refers to the process of translating a text from one language (source language) into another language (target language), using computers. The field of MT draws ideas and techniques mainly from linguistics, computer science, artificial intelligence (AI), translation theory and statistics [5]. In recent years, there has been a great increase in activity in the machine translation field, resulting in better translation quality being produced by MT systems. There are numerous motivations towards developing a computer based machine translator. First of all, this can help people perform translation faster and at a lower cost compared to human translators. There is also the perceived need to translate large amount of data and this can be done in a relatively short space of time using MT. Many different MT approaches have been developed, such as rule-based, statistical-based and example-based approaches. However there is still no MT approach that is able to produce high quality translations for broader domain systems. In fact most of the successful MT systems are for a restricted domain, such as the Canadian METEO system [6].

B. Machine translation architectures

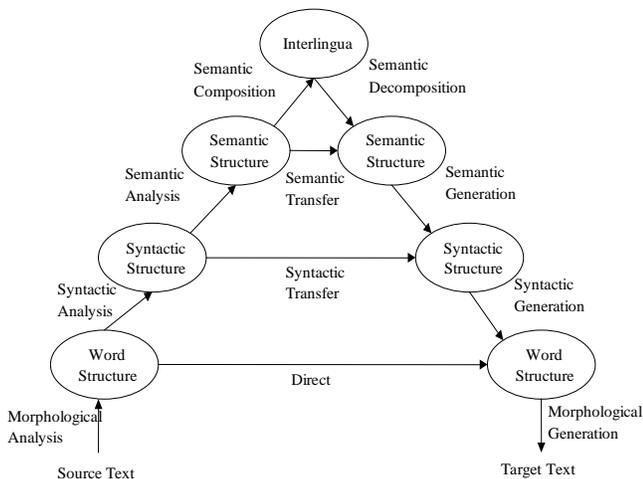


Fig.1. The Vauquois Triangle for MT [8]

Based on the level of linguistic analysis that is processed, MT system may be categorized as: direct, transfer, and Interlingua. The Vauquois triangle as shown in Figure 1 is used to illustrate these three levels. It shows the increasing depth of analysis needed as the top of the triangle is reached (i.e. from direct approach through transfer approach to Interlingua approach) [7]. Furthermore, the amount of transfer knowledge required to traverse the gap between languages decreases as the top of the triangle is reached.

1) Direct Architecture

In direct MT system, morphological analysis is performed on the source text to determine the core of the words to be translated. An example of this is the word “searching” would be analyzed as coming from the word “search”. After this stage, each word is translated into a specified language by using large bilingual dictionaries. Then the words are rearranged as according to the target language sentence format.

2) Transfer Architecture

The transfer-based MT architecture was developed to produce better translation quality by capturing and using the linguistic information of the source text. It consists of three stages. During the first stage, the source language (SL) text is analyzed to its syntactic structure (i.e. to produce a parse tree) or semantic structure (i.e. to determine the logical form). Then transfer rules are used to map the SL syntactic/semantic structure to a structure in the target language (TL). Finally in the third stage, the target text is generated from the TL structure [8].

3) Interlingua Architecture

In Interlingua MT system, the source text is analyzed and translated into a language-independent representation known as an Interlingua [9]. The target text is then generated from that Interlingua representation. A common choice of Interlingua used by MT systems is Esperanto.

C. Machine translation paradigms

The main types of MT system are introduced in this section.

1) Rule Based Machine Translation

A Rule-based MT (RBMT) system relies on linguistic rules to perform translation between the source and target language. The rules which are defined in such a system are extensively used in the various processes which analyze an input text, such as during morphological, syntactic and semantic analysis [10].

Rule-based approach is one of the oldest machine translation paradigms that have been developed. RBMT includes concepts such as transfer-based MT and interlingua-based MT. During the translation process in RBMT, the source text is analyzed to produce an intermediate representation (i.e. a parse tree or some abstract representation). The target text is then generated from that intermediate representation [10].

2) Statistical Machine Translation

Statistical MT (SMT) is a MT paradigm in which large bilingual corpora (i.e. a set of translated texts in both the source and target language) are analyzed to produce a translation model, and also large amounts of monolingual text in the target language are used to produce a language model [11]. The MT system then uses the two models to perform translation. The statistical approach has gained a growing interest in recent years, since its re-introduction by a group of IBM researchers in the early nineties. The idea was first proposed by Warren Weaver in 1949, but efforts in this

direction were abandoned for various philosophical and theoretical reasons [12].

3) Example-based Machine Translation

Example-Based MT (EBMT) was first suggested by Nagao in 1984 [13]. The basic idea behind EBMT is to generate translation based on previous translation examples. Like statistical MT, example-based MT uses a large bilingual corpus; from which large number of examples are extracted and stored in a database EBMT has three main stages [14]: (i) first of all, the source text is decomposed, and the resulting fragments are matched against a database of real examples, (ii) during the second stage, each fragment is translated and (iii) finally the target text is generated by recombining each fragment.

4) Hybrid Machine Translation

Hybrid approaches to MT integrates various MT paradigms, in an effort to make the most of the strengths of the individual paradigms, while compensating for their weaknesses. The basic idea behind hybrid approaches is to combine linguistic paradigms (i.e. RBMT) with non-linguistic paradigms, such as SMT or EBMT, to produce better results.

D. Mauritian Creole Grammar

The Mauritian Creole (MC) determiner system is quite different from that of French, and it can be considered to be much simpler, as there are no French definite and partitive articles. There is also the exclusion of grammatical gender as well as number in MC.

The core of the MC determiner system has the following functional elements:

- An indefinite singular article **enn** [15].
- A demonstrative **sa**, which is generally used in conjunction with **la** [15].
- A post-nominal specificity marker **la** [15].
- A plural marker **bann** [15].
- The morpheme **li**, which is used to represent the pronoun he/she/it/him/her, depending upon the context it is used.

Mauritian Creole verbs use TMA (Tense, Modality, and Aspect) markers to indicate the tense [16]. The tense marker 'ti' indicates an action that has already taken place (i.e. past tense). The modality marker 'pu' indicates something will happen (i.e. definite future) whereas the modality marker 'ava' is used to express something that may possibly happen (i.e. indefinite future). The aspect marker 'pe/ape' marks an action that is still going on (i.e. progressive), in contrast to the aspect marker 'finn/inn' which indicates an action that is already over (i.e. perfect) [17].

III. IMPLEMENTATION

The proposed system follows the following steps:

1) Split a text into an array of sentences using “.”, “!”, “?” as delimiters

2) Split each sentence into an array of words using “\W” meta sequence, and the underscore character as delimiters

3) A Greedy algorithm is used to find the longest match for a given fragment, in the database

4) Perform morphological analysis to extract root of word, and check for corresponding translation, in case word has not been translated in step 3.

5) Reorder the words according to the target language sentence format

The rule based machine system relies on the use of bilingual dictionary to perform translation. We have used the Diksioner Morisien dictionary to build the bilingual dictionary in the database.

A. Greedy Algorithm for Natural Language Processing

The greedy algorithm is used to retrieve the target word(s) from the database and it works in following way: it starts at the first character in a sentence, and by traversing from left to right it attempts to find the longest match, based on the words in the database. When a fragment is found, a boundary is marked at the end of the longest match, and then the same searching process continues starting at the next character after the match. If a word is not found, the greedy algorithms remove that character, and then continue the searching process starting at the next character [18]. The steps to perform the greedy search are given below.

1. Initialize startingPos to 0
2. Initialize numElementsWordArray to number of elements in wordArray
3. Initialize fragment to 5
4. Create an array translatedWordArray to Store target word
5. Create an array wordCategoryArray to Store word category
6. WHILE starting position < numElementsWordArray
7. Initialize flag to 0
8. fragment = fragment -1
9. Initialize fragmentEndPos to startingPos + fragment
10. IF fragmentEndPos > numElementsWordArray THEN
11. fragmentEndPos = numElementsWordArray
12. ENDIF
13. Create an empty String greedyString to Store the fragment of words
14. FOR (k= startingPos; k < fragmentEndPos; increment k)
15. greedyString = greedyString + “ ” + wordArray [k]
16. ENDFOR
17. IF flag is 0, and target word and word category is retrieved from database where source word = greedyString THEN
18. Put target word in translatedWordArray
19. Put word category in wordCategoryArray
20. Set startingPos to fragmentEndPos
21. Set fragment to 5
22. Set flag to 1
23. ENDIF
24. IF flag is 0, and word not found THEN
25. CALL morphologicalAnalysis (greedyString)
26. Store the variables received from the morphologicalAnalysis function in a List with parameters (target word, word category)
27. Put target word in translatedWordArray
28. Put word category in wordCategoryArray
29. Set startingPos to fragmentEndPos

```
30.          Set fragment to 5
31.  ENDFIF
32.ENDWHILE
```

The above greedy algorithm starts the searching process by taking a fragment of maximum of four words starting from the beginning of a sentence. If the fragment is not found in the database, the last word of the fragment is removed, and the searching process continues. If a fragment is found, a boundary is marked at its end, and the searching process continues taking another fragment of a maximum of four words, starting from the boundary.

B. Morphological Analysis

In case a word is not translated after the Greedy search sub-process, it enters the Morphological analysis process, where rules are applied to the word to find its root, which is then searched in the database. An example of a morphological rule is: the suffix -ing is removed from the word walking to obtain its root (i.e. walk). The steps for the morphological analysis process are as follows:

```
1.  Get word
2.  Set String truncatedWord to word
3.  Create an empty String saveLastChar to Store the last
    characters of a word
4.  Create an empty String translatedWordString to Store
    translated word
5.  Create an empty String wordCategoryString to Store word
    category
6.  IF length of word > 3 THEN
7.      FOR num= 1 to 4
8.          Set truncatedChar to last character of
            truncatedWord
9.          Add truncatedChar to saveLastChar
10.         truncatedWord = truncatedWord – last character
11.         IF num = 1 THEN
12.             IF the reverse of String saveLastChar =
                suffix THEN
13.                 IF target word and word
                    category is retrieved from database where source
                    word = truncatedWord THEN
14.                     Add data to
                        translatedWordString (optional)
15.                     Add target word to
                        translatedWordString
16.                     Add data to
                        translatedWordString (optional)
17.                     Add word category to
                        wordCategoryString
18.                     Break
19.                 ENDFIF
20.             ENDFIF
21.         ENDFIF
22.         IF num = 2 THEN
23.             Repeat steps 12-20
24.         ENDFIF
25.         IF num = 3 THEN
26.             Repeat steps 12-20
27.         ENDFIF
28.         IF num = 4 THEN
29.             Repeat steps 12-19
30.             Add the reverse of String saveLastChar
                to truncatedWord
31.             Add the uppercase of String
                truncatedWord to translatedWordString
```

```
32.          Add data to wordCategoryString
33.          Break
34.          ENDFIF
35.      ENDFIF
36.  ENDFOR
37.ELSE
38.  Add the uppercase of String word to translatedWordString
39.  Add data to wordCategoryString
40.ENDIF
41.RETURN the variables (translatedWordString,
    wordCategoryString) in an Array
```

C. Reorder word

After the words are translated, they are rearranged according to the target language sentence format. The pseudocode for reordering words is given below:

```
1.  Get translatedWordArray and wordCategoryArray
2.  FOR EACH element in wordCategoryArray
3.      Initialize key to the key of the array element
4.      Initialize value to the value of the array element
5.      IF key is not equal to 0
6.          IF wordCategoryArray [key -1] is equal
            to adjective, and value equal to noun
7.              CALL swapArrayElement
                (translatedWordArray, key-1, key)
8.              CALL swapArrayElement
                (wordCategoryArray, key-1, key)
9.          ENDFIF
10.         ENDFIF
11.     ENDFOR EACH
12.     RETURN the variables (translatedWordArray,
        wordCategoryArray) in an Array
```

IV. EVALUATION

Our goal is to provide the most accurate translation; therefore, whenever new rules were added, a series of tests was carried out to make sure that it does not affect the quality of translation.

Table 1 presents some sample translations obtained when translating sentences from English to Mauritian Creole.

TABLE I. TRANSLATION OF ENGLISH SENTENCES TO MAURITIAN CREOLE

Source text	Expected Result	Target Text
She is a brilliant student	Li enn zelev intelizan	Li ene zelev intelizan
I love spicy food	Mo kontan manze epise	Mo kontan manze epise
I can't tell if he is listening to me or not	Mo pa capav dir si lip ekoute mwa oubien non	Mo pa capav dir si li pe ekoute mwa oubien pa
Either take it or leave it	Swa pran li oubien les li	Swa pran li oubien dekale li

From the above table, we have shown that the translation obtained is of acceptable quality. The column 'Expected result' represents the result obtained from a manual translator while 'Target Text' is the text generated from the program. However for some cases, for e.g. in the last sentence, the word "leave" is being translated to the word **dekale** in Mauritian Creole, which is being used in the wrong context. The Creole word **dekale** means 'put it somewhere else' or 'move it

slightly' in English. This will be remedied in our future work where context will be used to deal with polysymy, i.e., words that have different meanings.

Table 2 presents some sample translation obtained when translating from Mauritian Creole to English.

TABLE II. TRANSLATION OF MAURITIAN CREOLE SENTENCES TO ENGLISH

Source text	Expected Result	Target Text
Done to disan, sauve ene lavi!	Give your blood, save a life!	Give your blood, save a life!
Apel mwa kan ou rente lakaz	Call me when you get home	Call me when you enter house
Dan dimans nou mete promotion lor diri	On Sunday we offer discount on rice	On sunday we put promotion on rice
Divin bon pou lasante	Wine is good for health	Wine good for health

As can be seen, most of the generated text is similar to those obtained by the human translator. In the last example, the word 'is' is missing as it is currently quite difficult to know when to add these types of words when translating from English to Creole. The rules are quite complex and there are too many exceptions. It is also challenging to deal with cases of synonymy, i.e. one word in the English language can be translated to several words with completely different meanings in Creole.

A. Weaknesses of the current system

Word sense disambiguation (WSD) is the process of identifying the appropriate sense of a word in a given context. This has not been addressed. Another weakness is that the system can deal with only short sentences. Thus, in our future work, we intend to improve the translation for longer sentences as well. The tool can also be converted into a web application so that it is easily accessible to everyone.

V. CONCLUSION

In this paper, we have implemented the first automated translation system that performs translation of English sentences to Mauritian Creole and vice-versa. The translation system uses the rule-based machine translation approach to perform translation. The results obtained show that the implemented system can provide translation of acceptable quality. The speed of translation of the system is also satisfactory. As part of our future work, we plan to investigate how the problem of word sense disambiguation can be solved and how translation can be improved for longer sentences. The

translation system would benefit both foreigners and the Mauritian population as it would enable them to swap between their mother tongue and Mauritian Creole with ease and convenience.

REFERENCES

- [1] QUIRIN, S., 2012. Kreol Morisien au Primaire. Week End, 15 Jan. p17.
- [2] CHAROENPORNSAWAT, P. AND SORNLERLAMVANICH, V. AND CHAROENPORN, T., 2002. Improving translation quality of rule-based machine translation. 19th International Conference on Computational Linguistics.
- [3] OLIVEIRA, F. AND WONG, F. AND HONG, I., 2010. Systematic Processing of Long Sentences in Rule Based Portuguese-Chinese Machine Translation. 11th Annual Conference on Computational Linguistics and Intelligent Text Processing, 21-27 March 2010, Iasi, Romania. Heidelberg: Springer, pp. 417-426.
- [4] POORNIMA, C. AND DHANALAKSHMI, V. AND ANAND, K.M. AND SOMAN, K.P., 2011. Rule based Sentence Simplification for English to Tamil Machine Translation System. International Journal of Computer Applications, 25 (8), 38-42.
- [5] HUTCHINS, W.J. AND SOMERS, H.L., 1992. An Introduction to Machine Translation. London: Academic Press.
- [6] CHANDIOUX, J., 1976. METEO, an operational system for the translation of public weather forecasts. Seminar on Machine Translation, 8-9 March 1976, Rosslyn, Virginia. Stroudsburg: Association for Computational Linguistics, 27-36.
- [7] JURAFSKY, D. AND MARTIN, J.H., 2009. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. 2nd ed. New Jersey: Prentice Hall.
- [8] NIE, J-Y., 2010. Cross-Language Information Retrieval. San Rafael, California: Morgan & Claypool Publishers.
- [9] MISHRA, R.B., 2011. Artificial Intelligence PB. New Delhi: PHI Learning Private Limited.
- [10] CARL, M. AND WAY, A., 2003. Recent Advances in Example-based Machine Translation. Heidelberg: Springer.
- [11] UEFFING, N. AND HAFFARI, G. AND SARKAR, A., 2007. Semi-supervised learning for Machine Translation. Machine Translation, 21 (2), 77-94.
- [12] ARMSTRONG, A-W. AND ARMSTRONG, S., 1994. Using large corpora. Cambridge: MIT Press.
- [13] NAGAO, M., 1984. A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. Artificial and human intelligence. Elsevier Science Publishers, pp. 173-180.
- [14] Somers, H., 2003. An overview of EBMT. In: M. CARL AND A. WAY, ed. Recent Advances in Example-based Machine Translation. Heidelberg: Springer, 3-57.
- [15] GUILLEMIN, D., 2011. The Syntax and Semantics of a Determiner System: A Case Study of Mauritian Creole. Amsterdam: John Benjamins Publishing Company.
- [16] ADONE, D., 1994. The acquisition of Mauritian Creole. Amsterdam: John Benjamins Publishing Company.
- [17] DICK, J-Y. AND AH-VEE, A. AND COLLEN, L., 2011. A Few Points on Verbs in Mauritian Kreol and in Creole Languages [online]. Port-Louis, Ledikasyon pu Travayer.
- [18] PALMER, David D., 1997. A trainable rule-based algorithm for word segmentation. Proceedings of the 35th annual meeting on Association for Computational Linguistics, 7-12 July 2007, Madrid, Spain.