American Journal of Engineering Research (AJER)
e-ISSN: 2320-0847 p-ISSN : 2320-0936
Volume-7, Issue-6, pp-170-176
www.ajer.org
Research Paper Open Access

# Lexical Ambiguity Resolution System For Standard Yorùbá Verbs

Adegoke-Elijah Adenike.[1], Odejobi Odetunji. A.[1],Salawu Akeem. S.[2]
[1](Department of Computer Science and Engineering, Obafemi Awolowo University, Nigeria)
[2](Department of African Languages and Linguistics, Obafemi Awolowo University, Nigeria)
**Corresponding Author: Adegoke-Elijah Adenike, aadegokeelijah@gmail.com

***ABSTRACT :****Machine Translation (MT) systems often face challenges in choosing appropriate translations for words in the source language. To resolve this problem, wordsense disambiguation systems depend on knowledge gained from different sources. The knowledge gained could be from external resources or corpora. Despite the importance of this knowledge sources, many world languages lack adequate training data required to undertake the task of resolving ambiguity that erupts when translating texts from one language to the other. Yorùbá language falls within this group. The objective of this study is to propose a method of resolving translation ambiguity in a Yorùbá to English MT system using limited resources. This research depends on a rule based method that utilizes the semantic feature of the predicate-object relation in Yorùbá sentences for the disambiguation. To evaluate the proposed method, twenty words with more than one possible English translation were tested on a Yoruba to English machine system. The result of the evaluation shows a precision of 85% and recall of 75%. This result shows that the use of selectional restriction is effective in resolving translation disambiguity in a Yorùbá-English machine translation system.*
***KEYWORDS*** *–Ambiguity, disambiguation, lexical, resolution, Yoruba*

---

Date of Submission: 1-06-2018 Date of acceptance: 16-06-2018

---

## I. INTRODUCTION

Human language is generally ambiguous. Word sense disambiguation is the task of determining the meanings of words in a context using computational approaches and also differentiating between the senses of the word [1]. Machine Translation is the automated translation of text from one language to the other. The idea of Machine translation was first conceived in the late 1940s by [2]. Decades after conception of the idea, much progress was however not achieved in the task of Machine translation as expected. [3] investigated the challenges militating against the development of machine translation system, and described the occurrence of semantic ambiguity or double meaning as a necessary problem a machine would never be able to solve without a "universal encyclopedia". Word Sense Disambiguation is a computational task of determining the right sense of a multi-meaning word using knowledge gained from its context. [4] refers to the task of choosing the correct translation of an ambiguous word in a context, for machine translation as translation ambiguity.

In the context of machine translation, the task of WSD system is to choose the correct target language translation of a given source language word, when the target language offers more than one possible translation. In a Yorùbá to English machine translation (Y-E MT) system, given a Yorùbá expression *Y, Ade lo jẹ eba lanaa,* the task of the ambiguity resolution system can be broken down into two stages. At the first stage, all the possible translations of the ambiguous verb *jẹ* are identified from a sense inventory. Most sense disambiguity systems make use of external resources such as bilingual dictionaries to identify the possible translations of a word. The possible translations of *jẹ* are *eat, owe and win.* At the second stage of the ambiguity resolution, suitable classification method is used to choose the right translation of the ambiguous word. The suitability of the classification method depends on the features of the languages under consideration, the availability of feature extraction tools and the type of knowledge sources available for the disambiguation task.

Corpus based approaches depend on large body of texts, that could be described as a good representation of the language under consideration, and makes use of machine learning algorithms to identify appropriate sense of an ambiguous word. Supervised models have been shown to consistently outperform knowledge- based models in all standard benchmarks [5], they however depend on sense-annotated corpora which are highly expensive and difficult to build [6]. Apart from the shortage of training data, a crucial

limitation of current supervised approaches is that a dedicated classifier (word expert) needs to be trained for every target word, making them less flexible, and hampering their use within end-to-end applications [7].

In contrast, knowledge based systems depend on the structural properties of lexico-semantic resources such as machine readable dictionary and ontology, and do not sense-annotated data [8]. Such systems construct models based on the underline resources, which is able to handle multiple target words at the same time, and disambiguate them jointly, whereas word experts are forced to treat each disambiguation target in isolation.

Despite the level attention that have been given to many languages such as English and French, with overwhelming amount of language resources, the development of sense disambiguity system for under-resourced languages has not been given the needed attention. The Yoruba language, even though has a great number of speakers, falls within the class of languages, for which more researches need to be carried out to improve the accuracy of applications such as information retrieval systems and machine translation systems. Although [9]developed an English to Yoruba machine translator for modified and not modified simple sentences, using phrase structure grammar and re-write rules, the developed system did not attempt to resolve ambiguity in any of the two languages under consideration. This study is aimed at studying the locus of ambiguity in Yoruba sentences, and proposing a method of resolving lexical ambiguity in the language, despite the unavailability of language resources tools.

Section II gives a brief description of Yorùbá language. Section III reviews some related literature in word sense disambiguation. Section IV discusses the method of resolving translation ambiguity in Yorùbá verbs. Section V discusses the results of our experimental evaluation. We conclude the study in section VI.

## II. YORÙBÁ LANGUAGE

Yorùbá Language is spoken majorly in the southwestern region of Nigeria, Togo, Brazil, Republic of Benn, Ghana, Sudan, Sierra-Leone and Cote D`ivoire. Outside Africa, a great number of speakers of the language are in Brazil, Cuba, including Trinidad and Tobago; the speakers of the language are estimated to be 30million.

Yorùbá sentences exhibit several forms of ambiguities. Lexical ambiguity occurs when an ambiguous word present in an expression renders the whole sentence ambiguous. For example, *Mo fúnAdéníOsàn*(I gave Ade Orange). The verb *fún*can be translated as *squeeze* or *give*.

Categorial ambiguity is a form a lexical ambiguity where the ambiguous word has meanings that belong to different grammatical classes.For example, *Tádé fi ata jẹ isu*(Tade ate yam with stew). The word *fi* can be translated as *put* (verb) or *with*(preposition).

Structural ambiguity occurs when ambiguity occurs as a result of the syntax or arrangement of the words in a sentence. For example, *Bàbáonílémèjì* (Two landlords). This expression can be translated *as a man with two houses* or *two landlords*.

Attachment ambiguity is a canonical form of structural ambiguity that is formed due to the prepositional phrase in the sentence. In this case, the prepositional phrase could be attached to either the subject noun phrase or the object noun phrase. For example, *Mo ríekuníorígi*(I saw a rat on the tree). The sentence does not answer in specific term the question, *who is on the tree?* The *rat* or *me*?

Semantic ambiguity occurs when a sentence can be interpreted in more than one way, even when the structural ambiguity or the lexical ambiguity present in the sentence have been resolved. For example, *Ó san ara* (She is fat). The sentence can actually be interpreted as *she washes her body* or *she is fat*.

This research work is focused on resolving lexical ambiguity that occurs when translating a verb in Yorùbá language text to the corresponding English Language text. For a number of verbs in Yorùbá language, there are more than one possible translation in English language. For example, the verb*kọ́* (high tone) can be translated as either build or *teach*, depending on the context (surrounding words). The verb gbá (high tone) can also be translated as either *sweep* or *kick*.

## III. RELATED WORKS

WSD is considered as AI-complete problem [10], that is, a task whose solution is as hard as most difficult problems in artificial intelligence; such as representation of common sense and encyclopaedic knowledge. The task can be described as an intermediate task, which is not an end to itself, but rather an essential task required to accomplish most natural language processing tasks. All disambiguation tasks involve matching the context of the instance of the word to be disambiguated with either information from an external source, or information about the contexts of previously disambiguated instances of the word derived from corpora [10]. Examples of external resources that can be used for knowledge driven WSD approach are machine readable dictionaries, thesauri, Wordnet etc. Knowledge driven approaches are Lesk Algorithm [11], Measurement of semantic similarity computed over semantic networks [12], selectional preference [13] and Heuristic methods [14]. Data driven approaches rely on knowledge derived from corpora. The corpora used may

be labeled corpora or unlabeled corpora. The systems that make used of labeled corpora are referred to as supervised based approaches [7], while those that relies on information derived from unlabeled corpora are known as unsupervised learning methods [15],[16].

All of the approaches described above require either external resources or textual corpora for training the disambiguation model. Prior to this research, there has been no previous known work on the Yorùbá verb sense resolution; this study is therefore aimed at developing a sense ambiguity resolution system with minimal resources to kick start research in this area of the language.

## IV. PATTERN OF RESOLVING LEXICAL AMBIGUITY IN YORÙBÁ VERBS

A SY sentence is made up of a noun phrase and a verb phrase [17]. Verbs are very important component of standard Yorùbá sentence, as no sentence can be meaningful without a verb [18]. A Yorùbá verb is found in a verb phrase, and is found immediately after a noun or noun phrase that acts as a subject in a given Sentence. One major characteristic of Yorùbá verbs is that they do not begin with a vowel, but rather with a consonant [4].

To understand the pattern of resolving ambiguity in Yorùbá-English machine translation system, a critical study of Yorùbá languages sentences (containing one ambiguous verb) translated to English language was done using two bilingual dictionaries [20] and [21].

If we consider the ambiguous verb *kọ́* in sentences a) and b) below:
a) *Adé kọ́ Bọ́lá* (Ade taught Bola)
b). *Adé kọ́ ilé* (Ade built a house)
It can be seen that the sense of the ambiguous verb *kọ́* is determined by the object of the verb. The verb *kọ́* translates to *teach* when the object is animate, and translates to *build* when the object is inanimate class.

If we consider another ambiguous verb *tọ́* in the sentences c) and d) below:
c) *Sadétọ́ èfọ́* (Sade picked vegetable)
d) *Gbénga tọ́ Adé* (Gbenga provoked Ade)

It can be observed that the sense of *tọ́*has nothing to do with the subject of the verb, but relies on the class of the object. The verb translates to *pick* when the object is *vegetable* and translates to *provoke* when the object is *animate,* and has nothing to do with the subject of the verb.

To further illustrate the point that the disambiguation of Yorùbá verbs depend mainly on its object, let us consider the ambiguous verb *bẹ̀rẹ̀.* This verb can be translated as *begin* or *squat.*
e) *Adé gíga ti bẹ̀rẹ̀*(Ade has began of Ade has squatted)
f) *Adé gíga ti bẹ̀rẹ̀ isẹ́* (The tall Ade has begun to work)

The disambiguation of the verb *bẹ̀rẹ̀* in sentence (e) appears impossible because it does not have an object. However, the ambiguity in the verb appears easier to resolve with the presence of *isẹ́* in (f), which suggests the *begin* sense of the verb. From the Illustrations above, we can deduce that the sense of these verbs depends on the class of the object of the verb. The sense classification method used in this study is selectional restriction. This is a rule based method that performs sense classification by imposing restriction on the class of object that a given verb can possess. This method is suitable because one of the attributes of Yorùbá verbs is their ability to enforce a selectional restriction on their possible subjects and objects [18].

From the deductions above, we realize the need to categorize possible Yorùbá nouns into classes. We have nine possible classes that the object of a given ambiguous verb can have. The classes are Liquid, Food, Animate, Inanimate, Vegetable, Properties, Places, Body Parts, Abstract and Music. Each of these classes has possible items which are indefinite. Class Liquid, for example, has possible items *omi* (water), *epo* (Palm Oil), *oti* (alcohol) etc.

**Table 1: Classes of Yorùbá nouns**

| Class | Items |
|---|---|
| Liquid | Omi (water), epo (palm oil), ọtí (alcohol) , obe (soup) etc. |
| Animate | Adé (name), Gbénga (name), Kìnìún (lion), Ajá (dog) etc. |
| Vegetable | Ẹ́fọ́(spinach), Ewédú (corchorus), Àlubọ́sá(Onions), Ata(pepper) etc. |

| | |
|---|---|
| Food | Ìresì (rice), Ẹ̀bà(cassava meal), Ogi(pap), Ẹran(meat), Ọbẹ̀(soup) etc. |
| Properties | Ilé(house), owó(money), ọkọ̀(car) |
| Inanimate | Ilé(house), Asọ(clothes). |
| Places | Oko(Farm), Ọjà(market), yàrá(room) |
| Abstract | Àlá(dream) |
| Body Parts | Apá(arm), ẹsẹ̀(leg), ojú(eye) |
| Musical Instrument | Ìlù(drum), agogo (bell) |

## V. LEXICAL AMBIGUITY RESOLUTION

To resolve the lexical ambiguity in Yorùbá monosyllabic verbs, the rule based approach of selectional restriction is used. This method becomes appropriate following our study of the pattern of resolving this ambiguity using bilingual dictionaries; the process of which is described above. When the system encountered an ambiguous verb, the succeeding object is extracted. The system extracts the semantic feature of the arguments of the verb by checking the class of the noun. The information of the class of the object, serve as main input for the disambiguation system. As an illustration, to disambiguate the verb *jẹ* in *Olú ló jẹ èbà náà*, the system checks out for the possible translations of *jẹ* from the machine readable bilingual dictionary. The word has three possible translations *eat, win and owe*. Next, the system extracts the object of the verb from the ambiguous sentence; *èbà* is extracted out. *Ẹ̀bà* belongs to the class food. This suggests to the system that the food sense of the verb *jẹ* is appropriate in this context. Thus, *jẹ* is translated as *eat*. Formal description of the process is as follow:

Y is a set containing tokens of a Yorùbá sentence

$Y = \{w_{-1}, w_0, w_1\}$ (1)

Where $w_{-1}$ is the subject of the ambiguous verb $w_0$, and $w_1$ is the object of $w_0$.

E is a set containing the possible translations of $w_0$.

$E = \{t_1, t_2, ..., t_n\}$ (2)

$t_i$ are the possible translations of $w_0$, where n is the total number of the possible translation of the verb.To choose the right translation for $w_0$, the proposed model depends on the hypernym( i.e. the class) of $w_1$. To derive the hypernymn of $w_1$, we loosely classified the nouns in Yorùbá language into nine categories.

$Y_n$is a set containing the possible classes of Yorùbá nouns.

$$Y_n = \{ C_l, C_a, C_v, C_f, C_{pr}, C_{in}, C_p, C_{ab}, C_{bp}, C_m \} \quad (3)$$

Where

$C_l$ is an infinite set containing all the elements of class*liquid*

$$C_l = \{Omi, epo, oti, obe, .....\} \quad (4)$$

$C_a$ is an infinite set containing all the elements of class animate

$$C_a = \{Adé , Gbénga , Kìnìún, Ajá, .....\}. \quad (5)$$

$C_v$ is an infinite set containing all the elements of class *vegetable*

$$C_v = \{Ẹ̀fọ́, Ewédú , Àlubọ́sá, Ata, .....\} \quad (6)$$

$C_f$ is an infinite set containing all the elements of class *food*

$$C_f = \{ Ìresì , Ẹ̀bà, Ogi, Ẹran, Ọbẹ̀, \ldots\} \quad (7)$$

$C_{pr}$ is an infinite set containing all the elements of class *properties*

$$C_{pr} = \{ Ilé, owó, ọkọ̀, \ldots\ldots\ldots\} \quad (8)$$

$C_{in}$ is an infinite set containing all the elements of class*inanimate objects*

$$C_{in} = \{ Ilé, Asọ\ldots\ldots\ldots..\} \quad (9)$$

$C_p$ is an infinite set containing all the elements of class *places*

$$C_p = \{ Oko, Ọjà, yàrá, \ldots\ldots\ldots..\} \quad (10)$$

$C_{ab}$ is an infinite set containing all the elements of class *abstract*

$C_{ab} = \{ Àlá, \ldots\ldots\ldots\ldots\}$ (11)

$C_{bp}$ is an infinite set containing all the elements of class *body parts*

$$C_{bp} = \{ Apá, ẹsẹ̀, ojú, ..\} \quad (12)$$

$C_m$ is an infinite set containing all the elements of class *musical instrument*

$$C_m = \{ Ìlù, agogo, .. \} \quad (13)$$

The class of is $w_1$ determined by checking the set to which it belongs.

Algorithm 1: Algorithm for the disambiguation method

```
Data: Yorùbá sentence
Result: Translated sentence
1 Yorùbásentence= Input Sentence;
2 sent= Parse(Yorùbásentence);
3 tokens= tokenize(sent);
4 [subject,verb,object]= extractSVO(tokens);
5 if verb in database then
6 if object in database then
7 att= fetchattribute (object);
8 trans(object);
9 else
10 att= msgbox("Enter class for object");
11 trans(object)= msgbox("Enter translation of  object");
12 end
13 trans(verb,att);
14 trans(subject);
15 Print translated sentence;
16 else
17 Print verb not in database;
18 end
```

**Explanation**

a. The *Parse* functionaccepts Yorùbá sentence.
b. The *Tokenize* function breaks down the parsed sentence into the constituent tokens.
c. *ExtractSVO* function mines the subject, verb and object from the tokens
d. *Fetchattribute* function gets the class of the object extracted in step c. from the database
e. *Trans* function translates the subject and object by simply looking up their translation from a bilingual dictionary, but translates the verb by using the class of the object.

## VI. SYSTEM EVALUATION

There are two major methods of evaluating WSD systems. *In vivo* evaluation evaluates the WSD system as a module embedded in applications. This method evaluates the system in so far as they contribute to the overall performance of a particular application, such as machine translator, information retrieval, or speech recognition. This approach is also referred to as adequacy evaluation. *In vitro* evaluation evaluates WSD systems independent of real life applications. Some of the measures employed in this evaluation method are *precision*, *coverage* and *recall* performance metrics.

To evaluate our system, we make use of both methods of evaluation. The in-vivo method used some selected ambiguous words in a machine translation system developed for this evaluation. Twenty monosyllabic verbs, with each not having less than two possible translations, were fed into the translator one at a time. Fig. 1 shows the translation of the ambiguous verb *tọ́* in the sentence *Mo tọ́ èfọ́*, using the machine translator. The verb can be translated into *pick* or *provoke*. The class of *èfọ́* determines the translation of *tọ́*. With the class of *èfọ́* been vegetable, the verb translates as pick, by using selectional restriction on the possible object of the verb. Out of the twenty monosyllabic verbs used for the testing, the proposed system performs satisfactorily in the task of resolving the lexical ambiguity by choosing the correct English translation for most of the ambiguous verbs. Our model was able to disambiguate seventeen (17) verbs, out of which fifteen (15) were correctly disambiguated. Using precision and recall performance metrics, the result represents precision of 85% and a recall of 75%.

## VII. RESULTS AND DISCUSSION

Despite its performance, the proposed method is not flexible. The non-flexibility of this method is due to the fact that each of the ambiguous verbs makes use of different hand-coded rules for its disambiguation. For the twenty verbs used in the machine translation system, a total of twenty different rules are therefore needed. For Yorùbá language that possesses vast number of ambiguous verbs, this method is therefore inadequate. The method discussed here did not make use of any external lexical resources such as machine readable dictionary or thesaurus. The absence of any of these tools may have been largely responsible for the non- flexibility of the method used. In future research, we hope to enhance the lexical ambiguity resolution system by incorporating an external knowledge resource for Yorùbá language.


**Figure 1: Evaluation of the model using In-vivo method.**

## VIII. CONCLUSION

In this paper, a system for choosing the most suitable translation of an ambiguous monosyllabic Yorùbá verb in a Yorùbá to English machine translation system was presented. The knowledge used in the resolution of this ambiguity is derived from the object of the verb. This knowledge is gained from identifying the class of the object of the verb. To identify the class of the object, we developed a database of possible noun classes and possible elements of each of the classes. The model proposed was implemented using Python programming language. This implemented system is useful and applicable for enhancing translation quality of a Yorùbá-English Machine Translation system. Experimental results indicate that the proposed method improves the accuracy of the machine translation system.

## REFERENCES

[1]. P. Borah, G. Talukdarand B. Arup, "Approaches for Word Sense Disambiguation-A survey", International Journal of recent Technology and Engineering (IJTERE), 3:1-4, 2014.
[2]. W. Weaver, "Machine Translation of Languages". In Locke, W. and Booth, D., editors, MachineTranslation of anguages, John Wiley & sons, New York. pp. 15-23. 1955.
[3]. Y. Bar-Hillel, "The present status of automatic translation of languages", Advances in computers 1,91-163, 1960.
[4]. G. Kikui, "Resolving translation ambiguity using non parallel bilingual corpora", Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language processing, 1999.
[5]. A. Raganato, J. Camacho-Collados, and R.Navigli. Word Sense Disambiguation: A Unified Evaluation Framework and EmpiricalComparison. In Proc. of EACL, pages 99–110, 2017
[6]. O. De Lacalle and E.Agirre.. A Methodology for WordSense Disambiguation at 90% based on large-scale Crowd Sourcing. In Proc. of SEM, pages 61–70, 2015.
[7]. T. Pasini and R. Navigli.. Train-O-Matic: Large-Scale Supervised Word Sense Disambiguation in Multiple Languages without Manual Training Data.In Proceeding of EMNLP, 2017.
[8]. D. Weissenborn, L. Hennig, F. Xu, and H. Uszkoreit. Multi-Objective Optimization for the Joint Disambiguation of Nouns and Named Entities. In Proc. of ACL, pages 596–605, 2015.
[9]. Safiriyu I. Eludiora, Odetunji A. Odejobi,"Development of an English to Yorùbá Machine Translator", International Journal of Modern Education and Computer Science (IJMECS), Vol.8, No.11, pp.8-19, 2016.DOI: .5815/ijmecs.2016.11.02
[10]. N. Ide and J. Veronis "Word sense disambiguation: The state of the art". Computational Linguistics, pages 1–40, 1998.
[11]. M. Lesk, "Automatic sense disambiguation using machine readable dictionary: How to tell a pine cone from an ice cream cone", Proceedings of the ACM-SIGDOC Conference, pages 24–26, Toronto, Canada, 1986.
[12]. C. Leacock, M. Chodorow, and G. Miller, "Using corpus statistics and Wordnet relations for sense Identification". Computational linguistics, pages147–165, 1998.
[13]. D. McCarthy. and J. Carroll. "Disambiguatingnouns, verbs and adjectives using automaticallyacquired selectional preferences", Computational Linguistics, pages 639–654, 2003.
[14]. [14] D. Martinez and E. Agirre. "One sense percollocation and genre/topic variations", Proceedings of the conference on Empirical Methods in natural language Processing (EMNLP), pages 207–215, Hong Kong, 2000.
[15]. H. Ng, B. Wang, and Y. Chan, "Exploiting paralleltexts for word sense disambiguation: An empiricalstudy", Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 455–462, 2003.
[16]. D.Yarowsky, Unsupervised word sense disambiguationrivalling supervised methods. In Proceedings of the 33rdAnnual Meeting of the Association for ComputationalLinguistic (ACL), pages 189–196, Cambridge, U.S.A, 1995.
[17]. L. O. Adewole. "Sequence and co-occurrence of Yorùbá auxiliary verbs". Canadian Journal of Linguistics, pages 1–1 7, 1989.
[18]. O. Awóbùlúyì, "The Yorùbá verb phrase. In Yorùbá Language and Literature", chapter 14, pages 225–246. University of Ife Press, 1982
[19]. A. Akinlabi,, "Facts about World's Languages: Encyclopaedia of the World's Languages, Past and Present" The H. W. Wilson Company, 2001.
[20]. L.O Adewole. "Bilingualized Dictionary of Yorùbá Monosyllabic Words."Montem Paperbacks, Akúré, Nigeria, 2005.

[21]. University Press Plc. “A Dictionary of the Yorùbá Language”, University Press, Ìbàdàn, 2011.

Adegoke-Elijah Adenike, Odejobi Odetunji A. ,Salawu Akeem S. "Lexical Ambiguity Resolution System For Standard Yorùbá Verbs.”American Journal Of Engineering Research (AJER), Vol. 7, No. 6, 2018, PP.170-176.