

Improving Adequacy in Rule Based English to Igala Automatic Translation System Through Word Sense Disambiguation

Abstract

The outputs of rule based English to Igala machine translation system: a system that translate English sentences to Igala language were evaluated with respect to adequacy. Adequacy is the measure of the quantity of information existent in the original text that the translated text contains; it indicates whether the output is a correct translation of the original sentence in a sense that the right meaning is transferred. Analysis revealed that adequacy was poor for sentences that contain words that exhibit ambiguity. The errors originated from the fact that correct meanings of ambiguous words were not selected based on the context. A further study of the architecture of the translation system showed it has not inbuilt module for word sense disambiguation, this accounts for the poor adequacy. A word sense disambiguation model was developed and incorporated into the MT system to disambiguate sentences before passing it to the translator. The model was implemented using JSP: a technology for building and deploying web applications in Java as the front end and MySQL as backend. The model was tested on 100 previously translated sentences that contain ambiguous words. The level of adequacy increased 32% to 69.6%.

Keywords: Translation, adequacy, rule based machine translation system, word sense disambiguation, ambiguity.

Introduction

Today, we live in an increasingly globalized and integrated world where lots of information is generated in various fields. However, since most of this information is in English, it remains out of reach of a large number of people for whom English is not the language of communication. As a consequence, an increasing demand for developing means of translating from one language to another to enable efficient communication across cultures is felt very strongly. Because human translation is expensive, time consuming and always in short supply, developing and deploying machine translation applications has become imperative for dealing with the problem of information inequalities created by multilingualism. Machine Translation is the use of computers to automate some or all of the process of translating from one language to another [1]. Research organizations and government agencies now develop tools for automatic translation of text in order to achieve wider outreach and bridge the gap of language diversity [8]. Igala is a language in Nigeria for which a significant percentage of the people cannot access information due to the fact information is mostly created in English. Developing Machine translation system that will efficiently translate English language to Igala to bridge this information gap has become a matter of absolute necessity in modern times.

Human language is ambiguous, so that many words can be interpreted in multiple ways depending on the context. Ambiguity occurs when a word can have two or more different meanings or senses. The table below illustrates this

Word	Senses
plant	living/factory
tank	vehicle/container

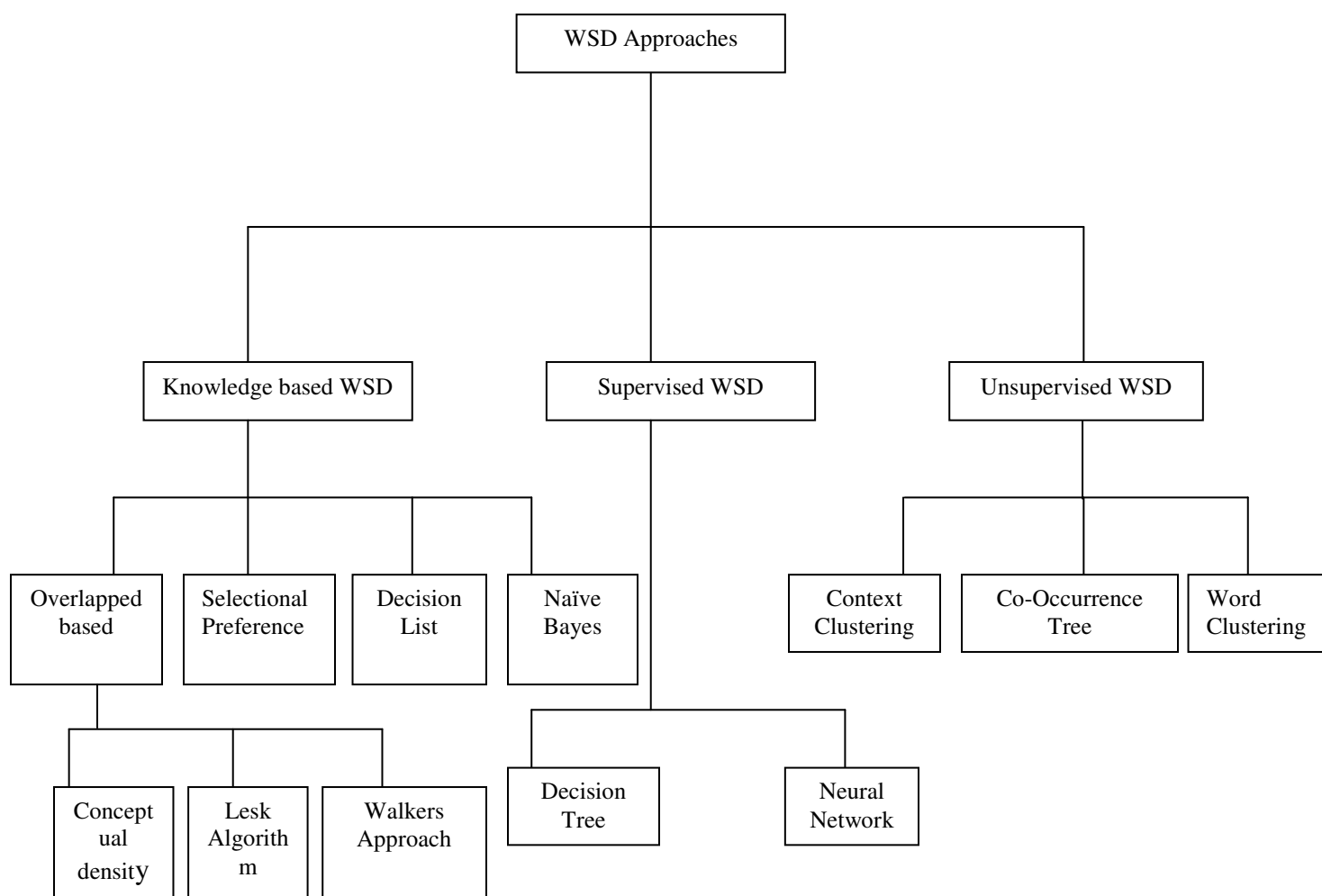
poach	steal/boil
palm	tree/hand
axes	grind/tools
sake	benefit/drink
bass	fish/music
space	volume/outer
motion	legal/physical
crane	bird/machine
execute	Perform/kill
cold	Disease/temperature

Ambiguity leads to translation ambiguity which occurs when a word in the source language can be rendered in more way in the target language.[4]. Selecting the right meaning of an ambiguous word when performing translation is critical to accuracy.

The task of assigning the correct or appropriate meaning to a given word in a text or discourse is called word sense disambiguation [7]. It is the computational identification of the meaning of words in context. The context of an ambiguous word is determined by the neighboring words. This is referred to as local or sentential context. In machine translation adequacy is the measure of the quantity of information existent in the original text that the translated text contains; it indicates whether the output is a correct translation of the original sentence in a sense that the right meaning is transferred. For adequacy to be achieved in machine translation, a word must be translated based on the context in which it is used. Word sense disambiguation therefore plays critical role in automatic translation of text.

The objective of this research is to model a language processor that can computationally determine the sense of an ambiguous word that is activated by its use in a particular context in a given English sentence before it is passed to the rule based English to Igala machine translator. The aim is to achieve significant improvement in translation quality.

A number of approaches have been proposed for WSD [5] [6]. Figure 1 is a diagram showing the different approaches. Overlapped based approach was adopted in developing our model.



Approaches to word sense disambiguation

Task description

Let $W(w_1, w_2, \dots, w_n)$ be the set of words left in the sentence after removing stop words. For each word w_i in W , verify that w_i is ambiguous by checking its sense indicator in a table. If the sense indicator of w_i shows that it is ambiguous, then there will be multiple senses S of w_i . The overlap for each sense of w_i is computed. The sense that has the maximum overlap is the most probable sense [3].

Algorithm

- i. Tokenize the sentence into word and store each token in array
- ii. Open stopwords table in database
- iii. Each token in array is searched for in the table, if found it is a stopword, it is removed from array. Repeat until all stopwords are removed from array.
- iv. An array of remaining words is left.
- v. Each token in the sentence is searched for in the lexicon table, if found the sense value is checked. If the sense value is “ps”, it is polysemous then

- vi. Open wsd table and retrieve first_sense, first_sense_relatedwords, second_sense, second_sense_relatedwords. Each sense related words are stored in separate arrays
First_sense_relatedwordsarray and second_sense_relatedwordsarray respectively.
- vii. Each word in First_sense_relatedwordsarray is compared with the words in remaining_wordsarray obtained in step iv. If words are found, the first_sense_overlap_number is incremented by 1. This process is repeated until all the first_sense_relatedwordsarray content are compared with remaining_wordsarray.
- viii. Each word in second_sense_relatedwordsarray is compared with the remaining_wordsarray, if a word is found the second_sense_overlap_number is incremented by 1. This process is repeated until all the second_sense_relatedwordsarray content are compared with remaining_wordsarray.
- ix. The overlap number for first_sense and second_sense are compared
If first_sense_overlap_number > second_sense_overlap_number, then first sense is selected as the most appropriate sense, otherwise the second sense is selected as the appropriate.
- x. The selected sense is used to replace the original token in the array.

Database Specification and Platform

The database for the system can be built on a number of database models. A database model is a type of model that determines the logical structure of a database and fundamentally determines in which manner data can be stored, organized and manipulated. There are basically four types of model: Hierarchical, Network, relational and object model. The most popular is the relational model which was introduced in by E.F Codd in 1970 [2]. A relational database model is one whose symbols are organized into a collection of relations. Relational model has a number of advantages over the other models such as ease of use, flexibility, precision, reduced data redundancy, reduced inconsistencies in records, data independence, easy addition and removal of data and increased security of data. Therefore the relational database model will be used for building the database and MYSQL is the relational database model of choice. The database schema is described below:

1. Table Name: wordsenses

s/n	Field Name	Data Type	Length
1	id	Auto number	6

2	ambiguous_word	VARCHAR	50
3	first_sense	VARCHAR	50
4	second_sense	VARCHAR	50
5	first_sense_relatedwords	VARCHAR	200
6	second_sense_relatedwords	VARCHAR	200

2. Table Name: multisensewords

s/n	Field Name	Data Type	Length
1	id	Auto number	6
2	word	VARCHAR	50
3	sense_indicator	VARCHAR	2

3. Table Name: stopwords

s/n	Field Name	Data Type	Length
1	id	Auto number	6
2	stopword	VARCHAR	50

System Architecture and Modules

Figure 2 is the architecture diagram of the overall of the proposed system. It has seven main components, gui, sanitizer, tokenizer, caseConverter, stopWordsRemover, disambiguator, assembler.

gui: creates the interface which is displayed for the user to enter the sentence to be disambiguated translated. The output of the disambiguation and translation is also presented to the user through this interface.

sanitizer: formats the input sentence. It removes all punctuation marks.

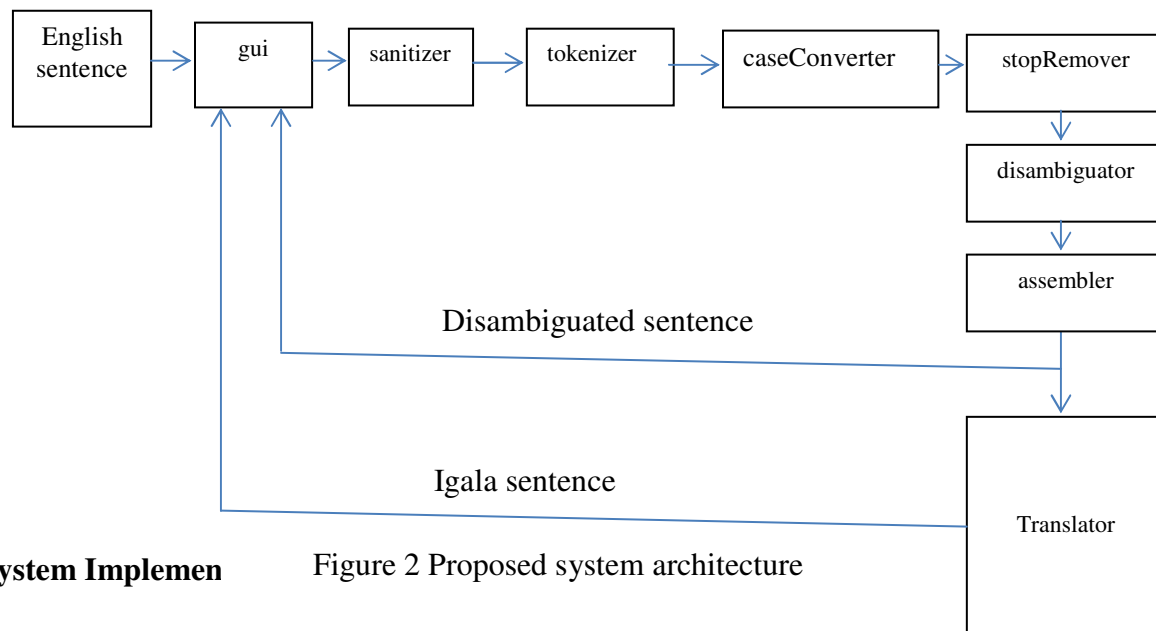
tokenizer: A basic text processing operation is tokenization which is the breaking up of raw text or sentence into words. This function is performed by the tokenizer. The input sentence is broken up at this point into words. It recognizes a word whenever a space is encountered which signifies the end of the word. The tokens are stored in an array.

caseConverter: converts all the tokens to lowercase.

stopWordsRemover: remove stop words from the array of tokens.

disambiguator: performs the task of selecting the correct sense for the ambiguous word present in the sentence. The original ambiguous word is replaced with the selected sense based on the context.

assembler: this recreates the sentence from the tokens. The sentence now contains the correct sense for the ambiguous word.



System Implemen

Figure 2 Proposed system architecture

The database tables were created in MySQL. The word sense disambiguation engine which applies a collection of computational rules to select the appropriate sense for an ambiguous word based on context was developed using JSP and integrated with the rule based English to Igala Automatic Translator which was also developed in JSP. The translation interface is shown in figure 3.

English to Igala Translator

Enter English Text to Translate
my watch showed a wrong time.
More
Disambiguated English Text
my clock showed a wrong time
Translated Igala Text
agogo mi du ña iko ebiene
Home

Figure 3. Translation interface

After entry of the English sentence, the user clicks the Translate button, the disambiguated sentence and the Igala equivalent is generated and displayed. The translate button caption changes to More as shown in the figure above.

Test and Evaluation

Translation quality is judged along two key dimensions namely; adequacy and fluency. Adequacy is the extent to which the meaning conveyed by the human translation is also conveyed by the machine translation being evaluated. It is the measure of the quantity of information existent in the original text that the translated text contains; it indicates whether the output is a correct translation of the original sentence in a sense that the meaning is transferred. Fluency on the other hand is the degree to which the translation is well-formed according to the grammar of the target language; fluency measures the extent of readability and understandability. The scale used for evaluating adequacy developed by Linguistics Data Consortium is the following: 5 All, 4 Most, 3 Much, 2 Little, 1 None [9].

100 sentences that were previously translated by the rule based English to Igala automatic translation System and found to have very low adequacy score due to ambiguity were retranslated after incorporating the word sense disambiguation model into the system. Previous good translations with the same ambiguous words were also retranslated as control; the outputs given were the same. The outputs were submitted for evaluation with respect to adequacy as it was done previously. Table 4 is a sample from the table of the scores.

AdequacyEvaluation					
sn	Englishsentence	PreviousMT	CurrentMT	PreviousAdequacyScore	CurrentAdequacyScores
1	my watch showed a wrong time.	mi go du ña iko ẹbiẹnẹ	agogo mi du ña iko ẹbiẹnẹ	2	4
2	The soldier executed the thieves.	choja lẹ kpa amoji lẹ	choja lẹ kpa amoji lẹ	5	5
3	The boy executed the job very well.	okolobia lẹ kpa ukọlọ lẹ nyọ nyọ gbali	okolobia lẹ che ukọlọ lẹ nyọ nyọ gbali	2	5
4	The headmaster gave present to the student who performed so well.	agboji skulu ẹdọ mẹfa lẹ du wa ńwu oma skul lẹ ki che nyọ nyọ gbali	agboji skulu ẹdọ mẹfa lẹ du ẹńwuojọ ńwu oma skul lẹ ki che nyọ nyọ gbali	2	5
5	I was present at the wedding ceremony.	U che wa icholo iyawo	U che wa icholo iyawo	5	5
6	Remember the role I played in your winning the election.	rewa ekwu U chiya efu wẹ ejẹ ijabe lẹ	rewa ekwu U che efu wẹ ejẹ ijabe lẹ	2	4
7	We will watch football in your house tomorrow.	awa á go ibolu efu unyi wẹ ọna	awa á go ibolu efu unyi wẹ ọna	5	5

Table 4

Previous percentage adequacy = 32%

Current percentage adequacy

$$= \frac{\sum \text{CurrentAdequacyScores}}{\sum \text{ExpectedAdequacyScores}} \times 100$$

$$= \frac{348}{500} \times 100$$

$$= 69.6\%$$

Conclusion

The paper concentrated on the design of a system that can computationally determine the sense of an ambiguous word that is activated by its use in a particular context in a given English sentence. The design was successfully implemented. A significant improvement in the level of adequacy was observed when the system was integrated with a rule based English to Igala automatic translator. Word sense disambiguation is therefore critical to achieving accuracy in automatic translation systems.

REFERENCES

1. Arnold, D, etal (1994) Machine Translation: An Introductory Guide, NCC Blackwell, London, ISBN: 1855542-17x.
2. Chip, D, Bob, B, Joseph, C.J, Mathew, W (2005) OCA: Oracle 10g Administration Study Guide, ISBN: 0-7821-4367-9, Sybex, San Francisco London.
3. J. Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amjan Shaik, P. Pavan Kumar, "Word Sense Disambiguation:
4. Marwan Akeel, R. B. Mishra, Divergence and Ambiguity Control in an English to Arabic Machine Translation. *Journal of Engineering Research and Applications* www.ijera.com ISSN : 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.1670-1679
5. Pranjal Protim Borah, Gitimoni Talukdar, Arup Baruah. *Approaches for Word Sense Disambiguation : A Survey*. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-3, Issue-1, March 2014.
6. Roberto Navigli. *Word Sense Disambiguation: A Survey*. ACM Computing Surveys, Vol. 41, No. 2, Article 10, Publication date: February 2009
7. Samit Kumar, Neetu Sharma, S. Niranjana. *Word sense disambiguation using Association Rules: A survey*. International Journal of Computer Technology and Electronics Engineering. ISSN 2249-6343. Volume 2, Issue 2. pp 93-98.
8. Sneha, Tripathi, Juran, Krishna Sarkhel, *Approaches to machine Translation* Anal of Library and information Studies, Vol. 57, December 2010, pp 388-393.
9. Xavier, G (2007) Survey of Machine Translation Evaluation, Universitat des Saarlandes, Computerlinguistik, Germany.