British Journal of Mathematics & Computer Science
11(1): 1-10, 2015, Article no.BJMCS.19994
ISSN: 2231-0851

SCIENCEDOMAIN *international*
www.sciencedomain.org

Improving Adequacy in a Rule-Based English-to-Igala Automatic Translation System through Word Sense Disambiguation

Sani Felix Ayegba[1*] and O. E. Osuagwu[2]

[1]Department of Computer Science, Federal Polytechnic Idah, Kogi State, Nigeria.
[2]Department of Computer Science, Imo State University, Owerri, Nigeria.

Article Information

DOI: 10.9734/BJMCS/2015/19994
Editor(s):
(1) Dariusz Jacek Jakóbczak, Chair of Computer Science and Management in this department, Technical University of Koszalin, Poland.
Reviewers:
(1) Anonymous, SASTRA University, India.
(2) Milam Aiken, University of Mississippi, USA.
(3) Ibrahim El-Zraigat, University of Jordan, Jordan.
Complete Peer review History: http://sciencedomain.org/review-history/10648

Original Research Article

Received: 06 July 2015
Accepted: 11 August 2015
Published: 23 August 2015

Abstract

Here, we describe a rule-based machine translation system that translates English sentences to the Igala language. This work was evaluated with respect to adequacy, a 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 communicated. An 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 no built-in module for word sense disambiguation. This accounts for the poor adequacy. A word-sense 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 a backend. The model was tested on 100 previously translated sentences that contain ambiguous words. The level of adequacy increased from 32% to 68.2%.

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

**Corresponding author: E-mail: felixsani@yahoo.com;*

1 Introduction

Today, we live in an increasingly globalized and integrated world where much information is generated in various fields. However, since most of this information is in English, it remains out of reach to a large number of people who do not speak the language. As a consequence, there is an increasing demand for developing a means of translating from one language to another to enable efficient communication across cultures. 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 in language diversity [2].

Igala is a language in Nigeria. A significant percentage of Igala speakers cannot access information due to the fact information is mostly created in English. Developing a 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. Here, we describe a rule-based English-to-Igala Machine Translation System developed to achieve the objective of efficiently translating between these languages automatically.

The human language is ambiguous because 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. Table 1 below illustrates this:

Table 1. Ambiguity illustrated

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

Translation ambiguity occurs when a word in the source language can be rendered in more than one way in the target language [3]. When performing translation it is necessary to select the right meaning of an ambiguous word, it ensures that the intended meaning in the source language is conveyed in the target language. The task of assigning the correct or appropriate meaning to a given word in a text or discourse is called ***word sense disambiguation*** [4]. 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 the sense that the right meaning is *communicated*. 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]. Fig. 1 is a diagram showing the different approaches. ***Overlapped-based*** approach was adopted in developing our model.

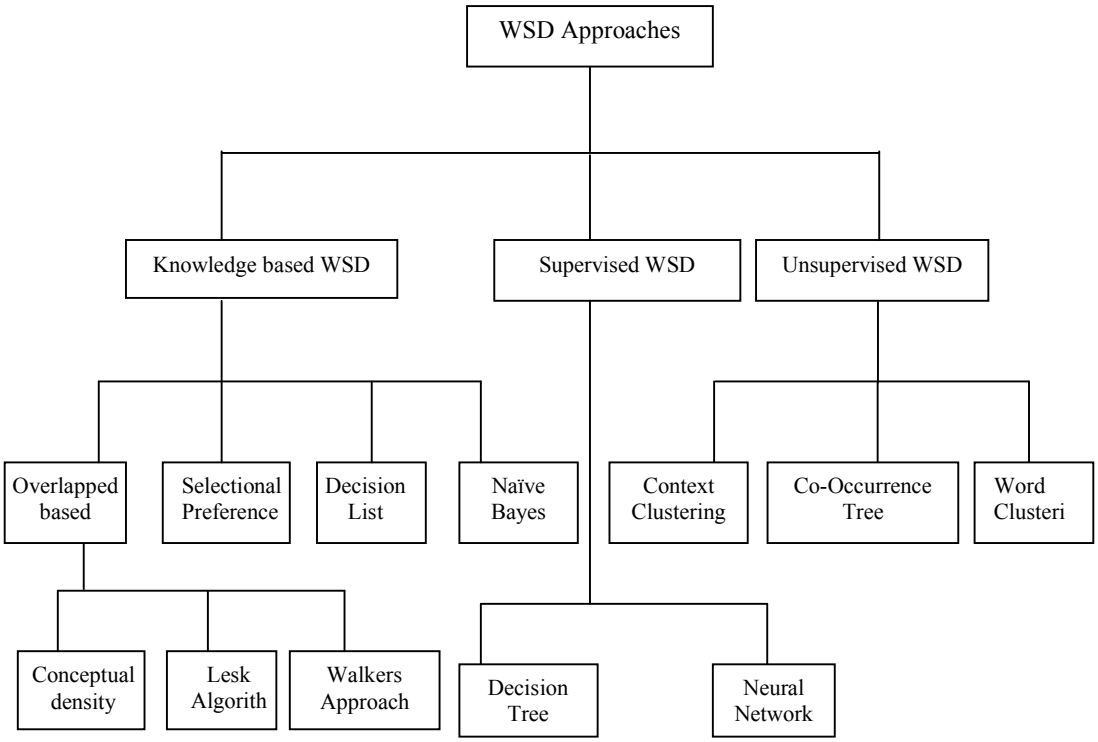

Fig. 1. Approaches to word sense disambiguation

1.1 Rationale for the Study

Absence of trained English to Igala language translators is making Igala speakers unable to participate in abundant business opportunities available in the online community and proper integration in the emerging Information society. Large proportions of the population do not speak or hear English that is a global language for business. Developing automatic machine translator for English to Igala language will be a major boost to economic activities in the territory of Igala nation.

2 Summary of Literature

[7] had posited that translation is critical for addressing information inequalities. A study conducted by Common Sense Advisory on behalf of Translators without Borders finds that translation is critical for the public health, political stability, and social wellbeing of African nations [8].

The work of translation was originally carried out by human translators. At a point the supply of translation services could no longer keep pace with the demand for translated content, moreover human translation is costly, time consuming and inadequate for addressing the real-time needs of businesses to serve multilingual prospects, partners and customers. The inherent limitations of human translation made the search for an alternative means of translation paramount. The search led to the discovery of what is known today as

machine translation or computer assisted translation. Machine Translation is the use of computers to automate some or all of the process of translating from one language to another [8].

Machine Translation systems (MT) can be classified according to their core methodology. Under this classification, two main paradigms can be found: The rule-based approach and the corpus-based approach. Within the rule based paradigm three approaches can be distinguished: Direct, Transfer and Interlingua [9] [10]. Rule-based systems are based on linguistically-informed foundations requiring extensive morphological, syntactic and semantic knowledge. Translation rules are created manually, demanding significant multilingual and linguistic expertise. Therefore, rule-based systems require large initial investment and maintenance for every language pair [11].

2.1 Limitations of Study

Machine Translators based on word sense disambiguation methodology has not been fully compared with other methods of automatic machine translation. Again the authors need to develop an online open access portal to critique the model for further refinement. We are unable to do this now until a proper refinement is achieved and the final version tested and attested for full functionality.

3 Methodology

The methodology adopted in this technical paper is ***overlapped-based model*** for Word Sense Disambiguation. A number of approaches have been proposed for WSD [5,6]. Fig. 1 is a diagram showing the different approaches.

3.1 Task Description

Let W (w_1, $w_{2...}w_n$) be the set of words left in the sentence after removing stop words. For each word $\mathbf{w_i}$ in W, verify that $\mathbf{w_i}$ is ambiguous by checking its sense indicator in a table. If the sense indicator of $\mathbf{w_i}$ shows that it is ambiguous, then there will be multiple senses S of $\mathbf{w_i}$. The overlap for each sense of $\mathbf{w_i}$ is computed. The sense that has the maximum overlap is the most probable sense [12].

3.2 Algorithm

i. Tokenize the sentence into word and store each token in an array
ii. Open stopwords table in database
iii. Each token in array is searched for in the table, if found to be a stopword, it is removed from the 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.

4 System Architecture and Modules

GUI (for graphic user interface) creates the interface which is displayed for the user to enter the sentence to be disambiguated before being translated. The output of the disambiguation and translation is also presented to the user through this interface.

The proposed System Architecture is presented in Fig. 2. It has seven main components – the GUI***, sanitizer, tokenizer, case Converter, stopWordsRemover, disambiguator, assembler***. ***The Sanitizer*** formats the input sentence and removes all punctuation marks. Tokenization 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.

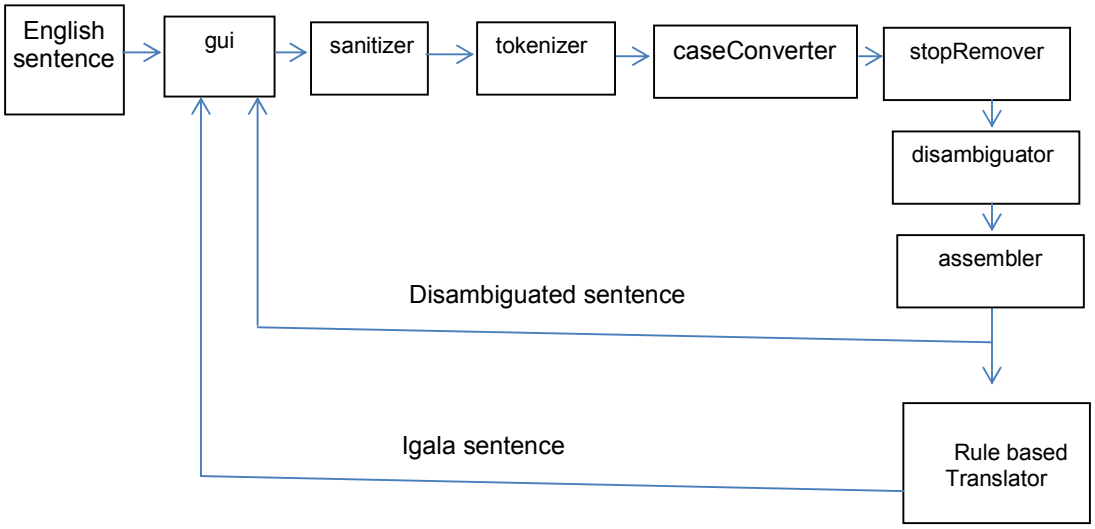

Fig. 2. Proposed system architecture

The ***Case Converter converts*** all the tokens to lowercase. ***StopWordsRemover*** removes 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. The ***Assembler*** recreates the sentence from the tokens. The sentence now contains the correct sense for the ambiguous word.

5 System Implementation

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 Fig. 4.

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

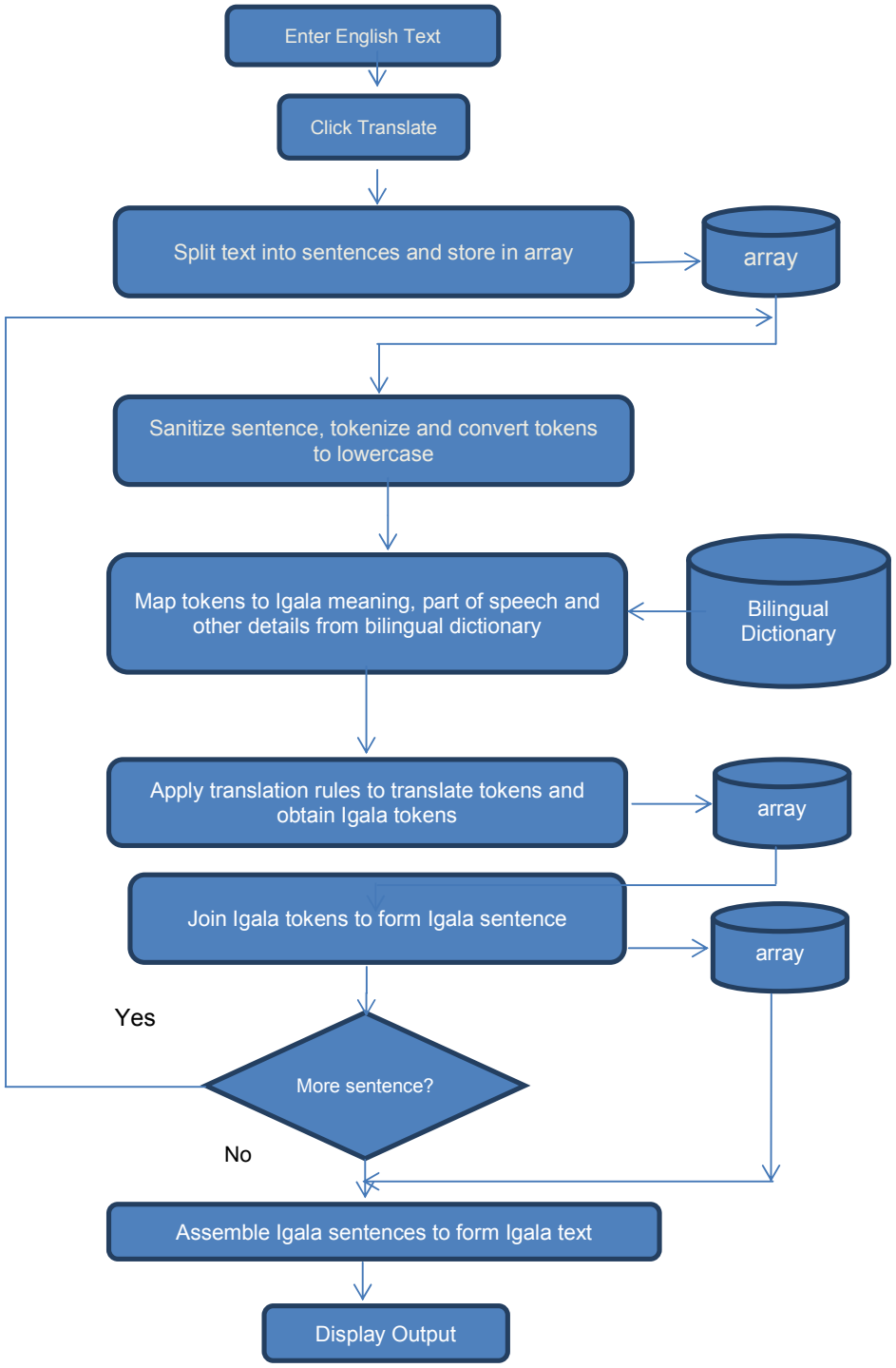

Fig. 3. Shows the steps in the translation of English to Igala by the Rule-based Translator component

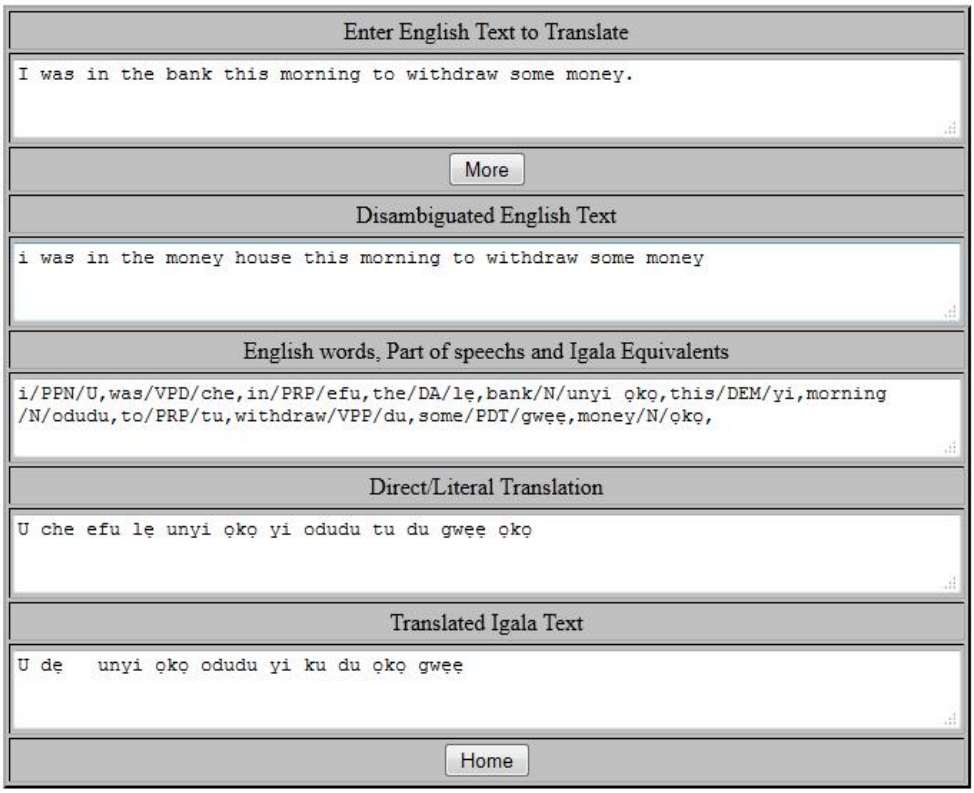

Fig. 4. Translation interface

5.1 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 output 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 and that used for fluency is: 5 Flawless, 4 Good, 3 Non native, 2 Disfluent, 1 Incomprehensible [13].

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. Two independent evaluators were used for the evaluation. One served as English to Igala translator for nineteen years at United Evangelical Church, Idah while the other is the staff of Radio Kogi, Ochaja, responsible for translating English documents to Igala before broadcasting in Igala language. Table 2 is a sample from the table of the scores.

Table 2. Adequacy and Fluency scores for translated sentences

	Adequacy Evaluation								
sn	**English sentence**	**PreviousMT**	**CurrentMT**	**PFScore**	**PAScore**	**CAScore1**	**CAScore2**	**FScore1**	**FScore2**
1	I went to the bank yesterday to withdraw money.	U le tu eti lẹ ọnalẹ ku du ọkọ	U le tu unyi ọkọ lẹ ọnalẹ ku du ọkọ	1	2	4	5	5	5
2	The soldier executed the thieves.	choja lẹ kpa amoji lẹ	choja lẹ kpa amoji lẹ	5	5	5	5	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	2	5	4	4	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 ẹñwuojo ńwu oma skul lẹ ki che nyọ nyọ gbali	2	2	5	5	5	5
5	I was present at the wedding ceremony.	U che wa icholo iyawo	U che wa icholo iyawo	5	5	5	5	4	4
6	Remember the role I played in your winning the election.	rewa ekwu U chiya efu wẹ eję ijabe lẹ	rewa ekwu U che efu wẹ eję ijabe lẹ	1	2	4	4	4	4
7	I will take my bath at the bank of the river.	U á gwẹ ọla mi eti aji lẹ	U á gwẹ ọla mi eti aji lẹ	4	5	5	4	5	4

PFScore (previous fluency score), PAScore (previous adequacy score), CAScore1 (Current Adequacy Score by first evaluator), CASore2 (Current Adequacy Score by second evaluator), FScore1 (Fluency Score by first evaluator), FScore2 (Fluency Score by second evaluator).

Average percentage adequacy obtained = 68.2%

6 Conclusion

This paper has 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. Any language translator that fails to embed WSD in the translation process will produce inaccurate translation. We therefore recommend its integration in all translation models for all languages.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Arnold D, et al. Machine translation: An introductory guide. NCC Blackwell, London, ISBN. 1994;1855542-17x.

[2] Sneha Tripathi, Juran Krishna Sarkhel. Approaches to machine translation anals of library and information studies. 2010;57:388-393.

[3] Marwan Akeel RB, Mishra. Divergence and ambiguity control in an english to arabic machine translation. Journal of Engineering Research and Applications. 2013;(3):6:1670-1679. ISSN: 2248-9622. Available:www.ijera.com

[4] Samit Kumar, Neetu Sharma, Niranjan S. Word sense disambiguation using association rules: A survey. International Journal of Computer Technology and Electronics Engineering. ISSN 2249-6343. 2012;2(2):93-98.

[5] Pranjal Protim Borah, Gitimoni Talukdar, Arup Baruah. Approaches for word sense disambiguation: A Survey. International Journal of Recent Technology and Engineering (IJRTE). 2014;(3):1. ISSN: 2277-3878.

[6] Roberto Navigli. Word sense disambiguation: A survey. ACM Computing Surveys. 2009;4(2):10.

[7] Sani Felix Ayegba, Osuagwu OE, Njoku Dominic Okechukwu. Machine translation of noun phrases from english to igala using the rule-based approach. West African Journal of Industrial & Academic Research. 2014;(11):1:17 -26.

[8] Ibid. 2014;17-26.

[9] Sneha Tripathi, Juran Krishna Sarkhel, Approaches to machine translation anals of library and information Studies. 2010;57:388-393.

[10] Omar Shirko, Nazlia Omar, Haslina Arshad, Mohammed Albared. Machine translation of noun phrases from Arabic to English using transfer-based approach. Journal of Computer Science. 2010;6(3):350-356. ISSN- 1549-3636.

[11] Hieu H. Improving statistical machine translation with linguistic; 2011.

[12] Sreedhar J, Viswanadha Raju S, Vinaya Babu A, Amjan Shaik P, Pavan Kumar. Word sense disambiguation: An empirical survey. International Journal of Soft Computing and Engineering. 2012;(2):2.

[13] Xavier G. Survey of Machine Translation Evaluation, Universitat des Saarlandes, Computer linguistik, Germany; 2007.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://sciencedomain.org/review-history/10648